



SAVONIA

PaTu

Patologian tutkimustietokannan suunnittelu

Henri Jokitörmä

Opinnäytetyö

Koulutusala Tekniikan ja liikenteen ala			
Koulutusohjelma Tietotekniikan koulutusohjelma			
Työn tekijä(t) Henri Jokitörmä			
Työn nimi PaTu, Patologian tutkimustietokannan suunnittelu.			
Päiväys	7.4.2013	Sivumäärä/Liitteet	55
Ohjaaja(t) lehtori Keijo Kuosmanen ja lehtori Veijo Pitkänen			
Toimeksiantaja/Yhteistyökumppani(t) Itä-Suomen yliopisto, tutkija Arto Mannermaa.			
<p>Tiivistelmä</p> <p>Tässä insinööriyössä tuotettiin ratkaisu Itä-Suomen yliopiston patologian yksikön rintasyöväntutkimusprojektin tiedontallennus ongelmaa. Projektissa useat tutkijat pitivät omia tutkimustiedostojaan omalla tietokoneellaan tallennettuina. Tämän lisäksi tiedostoja oli tallennettu yhteiselle verkkolevyllä. Tämä aiheutti ongelman silloin, kun haluttiin hakea viimeisintä versiota halutusta tutkimusaineistosta.</p> <p>Tämän ongelman ratkaisemiseksi suunniteltiin tietokanta ja kuvattiin tiedonsiirto lähdemateriaalista tietokantaan. Työ suunniteltiin MySQL Workbench -suunnitteluohjelmalla ja tiedonsiirtoa kuvannettiin Excel-taulukkoon. Työ tehtiin Itä-Suomen yliopiston tiloissa. Suunnitteluun käytettiin noin kolme kuukautta ja tästä ajasta käytettiin noin kuukausi tutkimusprojektiin perehtymiseen, sekä tarpeen kartoittamiseen.</p> <p>Työn tuloksena tuotettiin tämän raportin lisäksi MySQL Workbench:llä toteutettu suunnitelma tietokannasta ja tutkimusaineiston siirtoa tietokantaan kuvaavia dokumentteja, jotka eivät ole sisällytettyinä kokonaisuudessaan tähän dokumenttiin. Nämä tiedostot on luovutettu asiakkaalle.</p>			
Avainsanat Tietokanta, SQL, MySQL Workbench, SPSS, rintasyöpä, syöpätutkimus			

Field of Study Technology, Communication and Transport			
Degree Programme Degree Programme in Computer Science			
Author(s) Henri Jokitörmä			
Title of Thesis PaTu, Design of Database for Department of Pathology's Breast Cancer Treatment Project.			
Date	7 April 2013	Pages/Appendices	55
Supervisor(s) Mr. Keijo Kuosmanen, Lecturer and Mr. Veijo Pitkänen, Principal Lecturer			
Client Organisation/Partners University of Eastern Finland, Mr. Arto Mannermaa, researcher.			
<p>Abstract</p> <p>This thesis addressed the issue of data saving problem on a breast cancer treatment project in the department of pathology at University of Eastern Finland. Several researchers kept their own materials on their own computers along with some files being saved to shared network folders. This resulted in data fragmentation.</p> <p>To solve this data fragmentation problem a database was designed and dataflow between data source and database was mapped. MySQL Workbench was used to design the database and dataflow was outlined in Excel worksheets. This thesis was done in the facilities provided by University of Eastern Finland. The designing of this work took about three months from which about one month was used to getting familiar with the client's project.</p> <p>The result of this thesis included the database planned with MySQL Workbench, documents describing dataflow from source material to the database and this document. All these documents were delivered to the client.</p>			
<p>Keywords</p> <p>Database, SQL, MySQL Workbench, SPSS, breast cancer, cancer research</p>			

SISÄLTÖ

1	Johdanto	6
2	Solun muutos syöväksi.....	7
3	Syöpätutkimus.	9
4	Tässä työssä käytetyt työkalut ja tekniikat.....	12
4.1	Microsoft Excel 2010	12
4.2	MySQL Workbench.....	13
4.3	Microsoft SQL server 2008 r2	14
4.4	Relaatiotietokanta-malli	14
4.5	SPSS.....	15
5	Tietokanta	16
5.1	Tietokannan osa: Kohde	17
5.1.1	Kohde-taulun sisältö	18
5.1.2	Kohde-taulujen väliset suhteet.....	19
5.2	Tietokannan osa: Diagnoosi	20
5.2.1	Diagnoosi-taulun sisältö	21
5.3	Tietokannan osa: Kliiniset tiedot.....	22
5.3.1	Kliiniset tiedot-taulujen sisältö	23
5.3.2	Kliiniset tiedot-taulujen väliset suhteet.....	25
5.4	Tietokannan osa: Taustatiedot	27
5.4.1	Taustatiedon-taulujen sisältö	27
5.4.2	Taustatiedon-taulujen väliset suhteet	30
5.5	Tietokannan osa: Syväsekvenssi.....	31
5.5.1	Syväsekvenssi-taulujen sisältö	32
5.5.2	Syväsekvenssi-taulujen väliset suhteet	35
5.6	Tietokannan osa: RNA-ekspressio.....	37
5.6.1	RNA-ekspressio -taulujen sisältö	37
5.6.2	RNA-ekspressio –taulujen väliset suhteet.....	39
5.7	Tietokannan osa: Immunohistologia	42
5.7.1	Immunohistologia-taulujen sisältö.....	42
5.8	Tietokannan osa: Näytteet ja näytteiden sijainti.....	45
5.8.1	Näytteet-taulujen sisältö	47

5.9 Tietokannan osa: DNA snipit.....	50
5.9.1 DNA snipit taulujen sisältö	51
5.10 Tietokannan osa: Projekti	52
6 Ohjetiedostot.....	53
6.1 SPSS selite-tiedostot	53
6.1.1 SPSS selite-tiedostojen rakenne.....	53
7 Yhteenveto.....	55
Lähteet	56

1 Johdanto

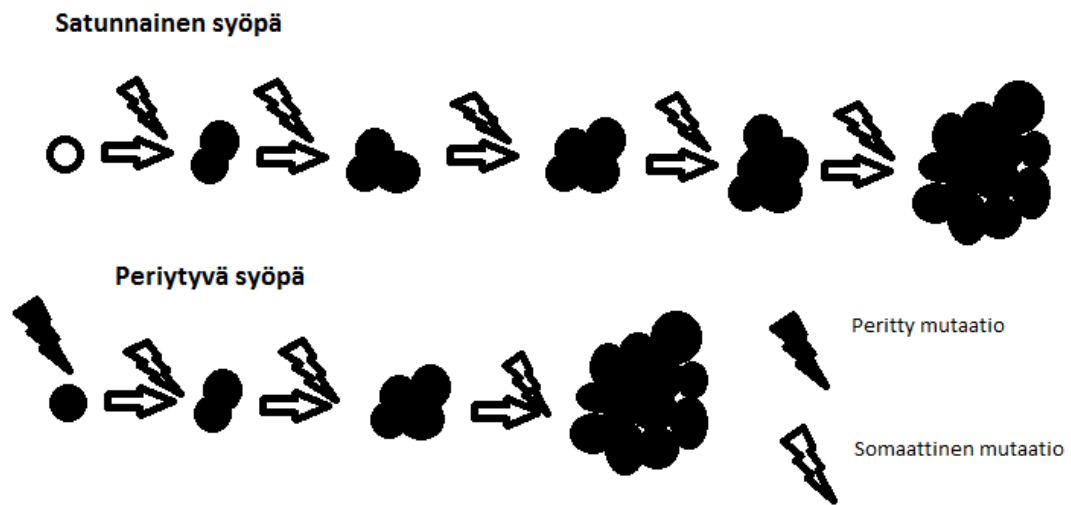
Tämän työn tilaaja on Itä-Suomen yliopiston lääketieteen laitoksen kliinisen lääketieteen yksikön patologian oppiaine, joka sijaitsee Kuopiossa yliopiston Canthia-rakennuksessa. Itä-Suomen yliopiston patologian yksikkö vastaa lääketieteen opiskelijoiden patologian koulutuksesta sekä patologian erikoislääkärikoulutuksesta. Tämän lisäksi patologian yksikössä tehdään tutkimustyötä, kuten rintasyövän tutkimusta, johon tämä työ liittyy. (Itä-Suomen yliopisto - Kliininen patologia ja oikeuslääketiede, 12.11.2012)

Tässä opinnäytetyössä aiheena on ratkaista Itä-Suomen yliopiston patologian yksikön rintasyöväntutkimusprojektin tiedontallennusongelmaa. Projektissa useat tutkijat pitivät omia tutkimustiedostojaan omalla tietokoneellaan tallennettuina. Tämän lisäksi tiedostoja oli tallennettu yhteiselle verkkolevylle. Tämä aiheutti ongelman silloin, kun haluttiin hakea viimeisintä versiota halutusta tutkimusaineistosta tiedostosta. Ratkaisuksi tähän ongelmaan suunniteltiin tietokanta, johon tallennetaan tietoja potilaista, terveistä verrokeista ja kaupallisista syöpäsolulinjoista sekä potilaiden ja verrokkien kliinisistä hoitotiedoista, taustatutkimustiedoista, verinäytteistä, kudospäätteistä, rna-ekspressio-, rna-syväsekvenssi- ja proteiinien värjäytymistiedoista.

Tavoitteena on saada kaikki tarpeellinen tieto yhteen paikkaan niin, että se olisi helposti kaikkien projektin tutkijoiden saatavilla. Lisäksi tietokannan suunnittelussa otettiin huomioon veri- ja kudospäätteiden säilytyspaikan kirjaaminen aina tiettyyn pakastimessa sijaitsevaan varastointilaatikkoon, jotta jo pakastinta avattaessa tiedettäisiin, mistä sieltä haluttu näyte löytyisi.

2 Solun muutos syöväksi.

Syöpä on geneettinen tauti. Syöpä ei tartu, mutta alttius sairastua sille voidaan saada perimässä. Kaikki syövät kehittyvät solussa tapahtuvien muutosten seurauksena, joko satunnaisesti usean somaattisen mutaation aiheuttaman muutoksen kautta tai perityn mutaation ja muutaman somaattisen mutaation kautta. Syöpä tarvitsee siis syntyäkseen aina yksilöllisiä mutaatioita (kuvio 1) (Aula, Kääriäinen ja Palotie 2006, 187 - 190).



KUVIO 1 Syövän kehitys (Aula, Kääriäinen ja Palotie 2006, 187 - 190)

Syöpää aiheuttavat DNA-mutaatiot voivat olla erikokoisia. Kun mutaatio osuu solujen jakautumista, kasvua ja erilaistumista tai ohjattua solukuolemaa sääteleviin geeneihin, ne alkavat muistuttaa syöpäsolua. Muutokset solussa ovat somaattisia eli yksilön elin-aikana syntyviä mutaatioita. Siten nämä muutokset eivät periydy. Periytyvän, korkean riskin syöpäalttiuden arvellaan liittyvän noin 5 – 10 %:iin kaikista syövistä. Näitä syö-välle vahvasti altistavia perimän muutoksia tunnetaan lähes kaikista syövistä. Niiden yleisyys vaihtelee eri syöpätyypeittäin, eikä kaikkia vielä tunneta (Aula, Kääriäinen ja Palotie 2006, 187 - 190).

Vaikka ympäristö- ja elämäntapatekijöillä on keskeinen osa useimpien syöpien synnyssä, on syöpä pohjimmiltaan geneettinen sairaus, joka syntyy solun perimään kertyvien vaurioiden tai mutaatioiden seurauksena. Toisin kuin normaaleissa kudoksissa, jotka ovat polyklonaalisia eli edustavat useaa eri solulinjaa, ovat syöpäsolut monoklo-naalisia, lähtöisin yhdestä emósolusta. Nämä syöviksi kehittyneet solut ovat saavutta-neet kasvuedun, kun muutama niiden avaingeeneistä on vioittunut (Aula, Kääriäinen ja Palotie 2006, 187 - 190).

Syöpäsolun tunnusmerkkejä ovat:

1. Omavaraisuus kasvua edistävästä signaaleista, syöpäsolu tuottaa itse omat kasvusignaalin
2. Riippumattomuus kasvua rajoittavista signaaleista, kasvua rajoittavat geenit ovat inaktivoituneet
3. Ohjelmoidun solukuoleman, aptoosin, väistäminen.
4. Rajaton kyky jakautua
5. Verisuonten uudismuodostuksen eli angiogeneesin ylläpito
6. Kyky tunkeutua ympäröivään kudokseen ja muodostaa etäpesäkkeitä

(Aula, Kääriäinen ja Palotie 2006, 187 - 190)

3 Syöpätutkimus.

Itä-Suomen yliopiston patologian yksikön tutkijaa Arto Mannermaata haastateltiin tätä työtä varten. Arto Mannermaa toimii tässä projektissa asiakkaana. Tämän haastattelun tarkoitus on johdattaa lukija tähän projektiin ja tuoda syöväntutkimusta lukijalle jonkin verran tutuksi. Haastattelu tehtiin 2.11.2012.

Tämän kaltaisia syöpätutkimusprojekteja tehdään paljon. Käytännössä kaikissa Suomen yliopistosairaaloissa on meneillään tämän tapaisia tutkimusprojekteja. Tämä on eettisen toimikunnan hyväksymä tutkimus, johon mukaan tullessaan potilas allekirjoittaa suostumuksen KYS:iin tullessaan. Tutkimuksessa jälkeen hänestä otetaan näytteet ja täytetään haastattelulomake. Jos hän ei halua osallistua tutkimukseen niin sitten hänestä ei oteta tähän tutkimukseen mitään näytteitä eikä hänen tietojiaan kerätä. Tutkittavalla on myös milloin tahansa oikeus jättäytyä pois tutkimuksesta.

Tällaiset tutkimukset ovat pitkiä. Tällä hetkellä käytetyin materiaalimme on kerätty vuosina 90 - 95, eli siitä alkaa olla 20 vuotta. Tämä materiaali on nyt kypsimmillään, koska tiedetään kuinka syöpä on näissä tutkittavissa potilaissa edennyt. Kuinka tappava se on ollut ja mitä syöpiä on pystytty hoitamaan niin, että tauti ei ole uusiutunut.

Tällä hetkellä kerättyjä uusia tietoja voidaan toki käyttää diagnostiikkaan eli taudin tyyppin määrittämiseen, mutta koska näillä ei ole tarvittavaa seuranta-aikaa niin taudin luonteeseen ei voi puuttua vielä ollenkaan. Kuitenkin tästä vanhemmasta aineistosta taudin luonnetta voidaan tutkia.

Näistä potilaista otetuista näytteistä tutkitaan laboratoriomenetelmin pääasiassa kolmea tasoa, eli DNA-, RNA- ja proteiini-tasoa. DNA-tasosta me tutkimme DNA:n muutoksia, kuten yhden emäksen muutoksia, ”snipejä” tai tehdään sekvenssi analyysi. RNA-tasossa tutkimme geenin määrää eli ekspressiota. Proteiineja tutkitaan taas esimerkiksi immunohistokemisesti parafiiniblokeista. Snipi-analyysijä tehdään täällä itse. Sekvenssianalyysijä teemme jonkun verran itse, mutta pääasiassa näyte lähetetään muihin laboratorioihin koko perimän sekvensointiin.

Kaikissa näissä tutkimuksissa yritetään löytää yhteys sen tutkimustuloksen ja jonkin syöpäsairauteen liittyvän asian, esimerkiksi syöpätyypin tai potilaan parantumisen välillä. Esimerkiksi jonkin proteiinin immunohistokemiallisen määrityksen tuloksia, kuten

proteiinin määrää, vertaillaan jonkun erityyppisissä kasvaimissa. Voidaan myös tutkia vaikuttaako proteiinin määrä kasvaimessa potilaan ennusteeseen jollakin tavalla. Pitkän seuranta-ajan jälkeen voidaan todeta kuinka potilaalle annetut hoidot ovat vaikuttaneet. Voidaan katsoa esimerkiksi vaikuttaako tietty proteiini kasvaimen aggressiivisuuteen jollain tavalla. Eli periaatteessa näitä tuloksia peilataan aina klinisiin potilasta koskeviin tietoihin ja taustatutkimustietoihin. Esimerkiksi. Tutkitaan kuinka potilaan ruokavalio, liikunta tai hormonaaliset tekijät kuten kuukautiskierron pituus ja kesto tai lasten lukumäärä vaikuttaa tautiin ja sen hoitoon. Siten sekä perimän vaihtelua että taustatietoja tutkitaan yhdessä. Sen sijaan emme tutki liittykö esimerkiksi jonkin proteiinin määrä kasvaimessa siihen kuinka paljon kaakaota potilas on elämänsä aikana juonut.

Immunohistokemiallisen tutkimuksen tekemiseen parafiininäytteille. Parafiini on materiaalia johon varsinainen kudos on säilöty. Ennen kuin kudos säilötään parafiiniin, se kivetetään formaliinilla. Formaliini tappaa kaikki entsyymit ja muut molekyylit, jotka tuhoavat normaalisti näytteen solurakennetta. Kivetetty näyte säilötään parafiiniin, jotta solurakenteet säilyisivät.

Immunohistokemia on menetelmä, jolla tutkitaan parafiininäytteen solurakenteita. Tähän käytetään teollisesti valmistettuja vasta-aineita jotka pystyvät tunnistamaan kiinnostuksen kohteena olevia proteiineja solussa. Parafiiniblokista leikataan hyvin ohut leike ja se värjätään vasta-aineen sisältävällä liuoksella. Vasta-aine sitoutuu tiettyyn proteiiniin. Mitä enemmän leikkeessä on ko. proteiinia, sitä enemmän leike värjäytyy. Leikettä tutkitaan mikroskoopin avulla. Värjäntyneiden solujen määrä ja värjäntymisen vahvuus kirjataan ylös, samoin proteiinin sijainti solussa.

DNA-snipit ovat DNA-muutoksia joita esiintyy jokaisella ihmisellä. Tässä tutkimuksessa niitä katsotaan normaaliperimästä eli veren valkosolujen DNA:sta. Josta voidaan tutkia jonkin geenimuodon liittymistä tutkittavaan tautiin, tässä tapauksessa rintasyöpään. Esimerkiksi tutkimme jonkin solussa sijaitsevan entsyymin geneettistä muunnosta, joka kiihdyttää entsyymin toimintaa. Haluamme siten selvittää altistaako lisääntynyt entsyymin toiminta rintasyövälle. Esimerkiksi entsyymi hajottaa maksassa syövänhoidossa käytettyä solumyrkkyä. Jos entsyymin toiminta on kiihtynyt, niin se voi heikentää hoidon tehoa ja täten heikentää potilaan ennustetta.

Näitä snippejä tunnetaan miljoonia ihmisen perimästä ja ne eivät yksinään vaikuta kovin paljoa syöpäriskiin, mutta jos niitä on vaikkapa 30 jotka kukin lisäävät hiukan riskiä sairastua syöpään niin silloin yhteisvaikutus voi olla jo aika merkittävä tekijä syövän kehittymiselle.

SNPit voi olla myös geeneissä joiden proteiinit osallistuvat DNA-vaurioiden korjaamiseen. Kun solussa tapahtuu jokin mutaatio, niin solu pyrkii korjaamaan mutaation käyttämällä mallina tervettä DNA-juostetta. Jos solussa on kuitenkin heikentynyt mekanismi vaurion korjaamiseen, niin solu ei aina pysty mutaatiota korjaamaan ja mutaatio voi jäädä soluun ja periytyä edelleen solun tuottamiin tytärsoluihin. Tytärsoluihin voi syntyä uusi mutaatio joka voi jäädä korjaamatta ja näin mutaatioita saattaa kertyä esimerkiksi kahdenkymmenen vuoden aikana niin paljon, että kasvain syntyy.

Verinäytteistä haluamme tutkia henkilön normaalia perimää. Nämä perintötekijät voivat vaikuttaa taudin kehittymiseen tai etenemiseen. Normaali perimää taas voidaan verrata syövän perimään, jolloin saadaan selvillä mahdolliset solussa tapahtuneet muutokset.

Kasvaimen tuorenäyte sisältää kaikki ne mutaatiot jotka ovat aiheuttaneet sen solun muuttumisen syöpäsoluksi. Sitä tutkitaan DNA-tasolla eli perimätasolla ja RNA-tasolla eli miten eri geenit ilmentyvät. Kasvaimissa on aktivoituneita syöpägeenien ilmentämiä proteiineja joita tuotetaan suuria määriä normaalitilanteeseen verrattuna ja jotka aiheuttavat syöpäsolun jatkuvan kasvun ja jakautumisen ja lopulta etäpesäkkeiden eli metastaasien muodostumisen muualle ihmisen kehoon. Paikallinen kasvain ei ole vielä välttämättä pahanlaatuinen syöpäkasvain. Se muuttuu pahanlaatuiseksi eli maligniksi kun se alkaa levittää syöpäsoluja ympäriinsä. Kasvain joka ei ole vielä lähettänyt etäpesäkkeitä on usein hoidettavissa.

Kaikki syöpätutkimukset tehdään tuorenäytteille. Paitsi immunohistokemialliset tutkimukset tehdään parafiininäytteille, jotka eivät enää ole tuorenäytteitä. Tietysti parafiininäytteet tehdään alun perin tuorenäytteistä. Tuorenäytteitä otetaan kasvaimen poiston yhteydessä.

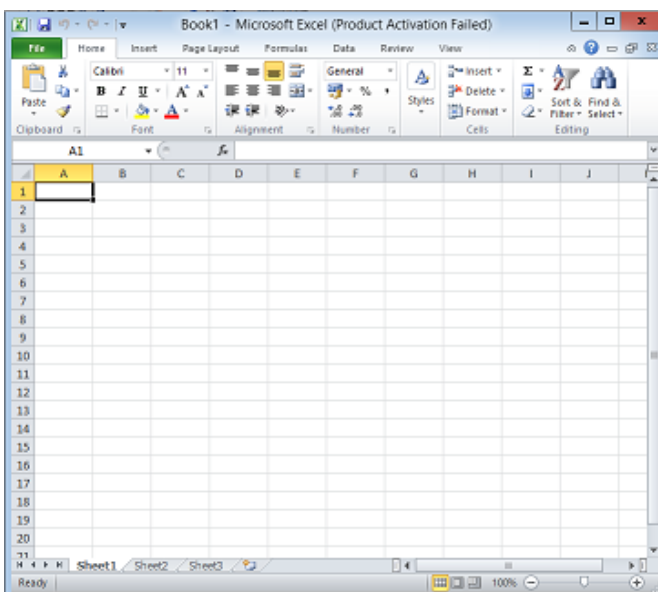
4 Tässä työssä käytetyt työkalut ja tekniikat

Tämä projekti toteutettiin työskentelemällä Itä-Suomen yliopiston patologian yksikönteiloissa. Näin oli helpompi perehtyä syöväntutkimukseen jota patologian yksikössä tehtiin.

Tietokannan suunnitteluun käytettiin MySQL Workbench 5.2 – työkalua. Tämän lisäksi luotiin ohjetaulukoita, kuvaamaan tiedonsiirtoa lähdemateriaalista tietokantaan, Microsoft Excel 2010 – ohjelmalla. Työn aikana toteutettiin muutamia demoja suunnitellun kannan toiminnasta. Tätä varten asennettiin VMwarella virtuaalikoneen johon asensin Microsoft SQL server 2008r2 – tietokannan. Tiedon siirtoon käytettiin C#-kielillä kirjoitettua ohjelmaa, joka luki ja muokkasi lähdemateriaalin csv-tiedostoista sql-lauseita, jotka siirsin virtuaalikoneella sijaitsevalle SQL serverille. Lähdemateriaali oli tallennettuna IBM:n SPSS ohjelmalla. Tätä työkalua käytettiin tässä työssä vain csv-tiedostojen tuottamiseen lähdemateriaalista.

4.1 Microsoft Excel 2010

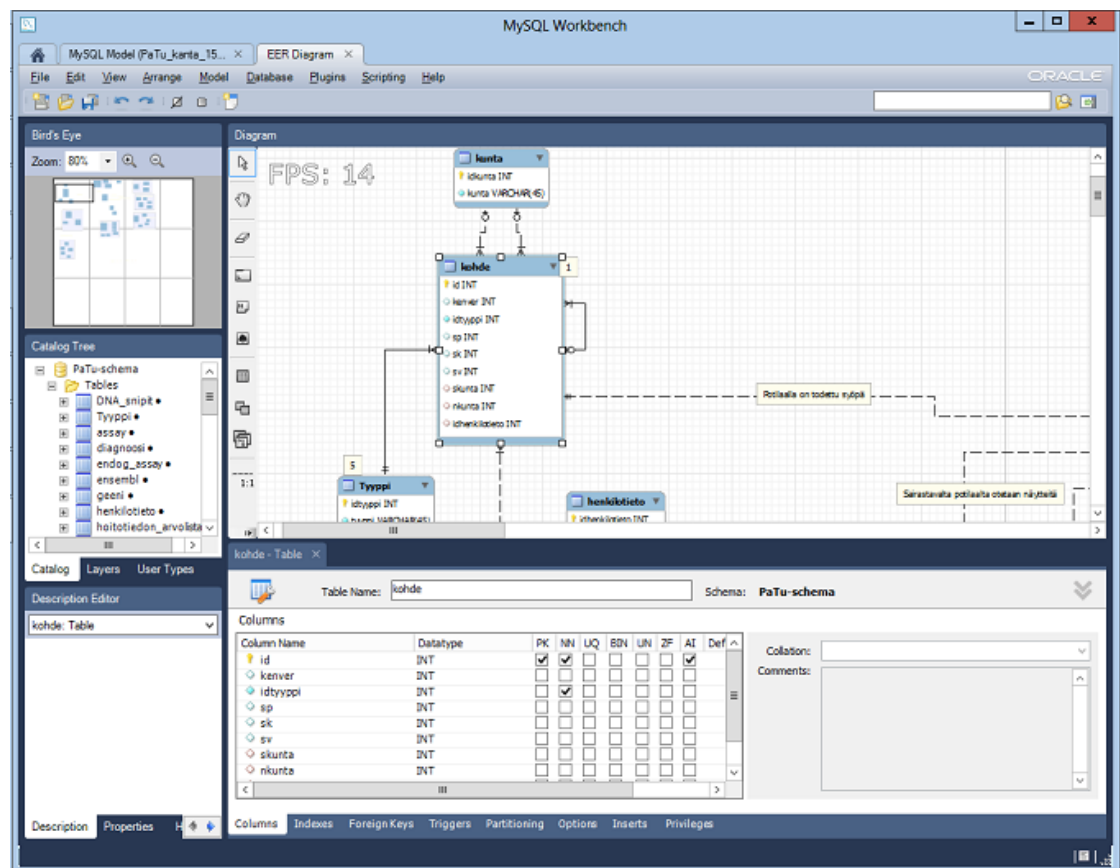
Microsoft Excel on Microsoftin alun perin vuonna 1985 Macintosh tietokoneille julkaissama taulukkolaskentaohjelma. Vuodesta 1987 lähtien Excel on ollut myös saatavilla Windows tietokoneille. Tällä kirjoitus hetkellä tuorein versio Microsoft Excelistä on 2010 (kuva 1). (Power, 14.11.2012)



KUVA 1 Microsoft Excel 2010

4.2 MySQL Workbench

MySQL Workbench (kuva 2) on Oraclen kehittämä visuaalinen SQL-tietokantojen kehitys-, ylläpito- ja suunnittelutyökalu. Kuten jo ohjelman nimestä voi päätellä, niin Workbench on suunniteltu erityisesti MySQL-tietokantoja varten, mutta soveltuu myös muiden SQL-kantojen suunnitteluun. Tällöin kuitenkin ohjelman tuottamat tietokannan luontilauseet eivät ole täysin yhteensopivia ja niitä joutuu hieman muokkaamaan jotta ne soveltuisivat käytettäväksi muissa SQL-tietokannoissa. MySQL Workbench on ilmainen ohjelmisto joka on lisensoitu GNU General Public License – lisenssillä. (Oracle, 14.11.2012)



KUVA 2 MySQL Workbench

MySQL Workbench osoittautui varsin tehokkaaksi ja ketteräksi tietokantojen suunnittelu työkaluksi.

4.3 Microsoft SQL server 2008 r2

SQL server 2008r2 on Microsoftin kehittämä SQL-relaatiotietokantojen hallintajärjestelmä. Tämän työn teko hetkellä uusin versio SQL serveristä on 2012. Tähän työhön valittiin kuitenkin 2008r2 version koska se oli tutumpi työn tekohetkellä. (Microsoft, 14.11.2012)

Kuten aikaisemmin mainittiin (kappaleessa 2) SQL serveriä käytettiin vain suunnitellun tietokannan demoamiseen. Toteutus voidaan kuitenkin tehdä käytännössä millä tahansa SQL-tietokannalla.

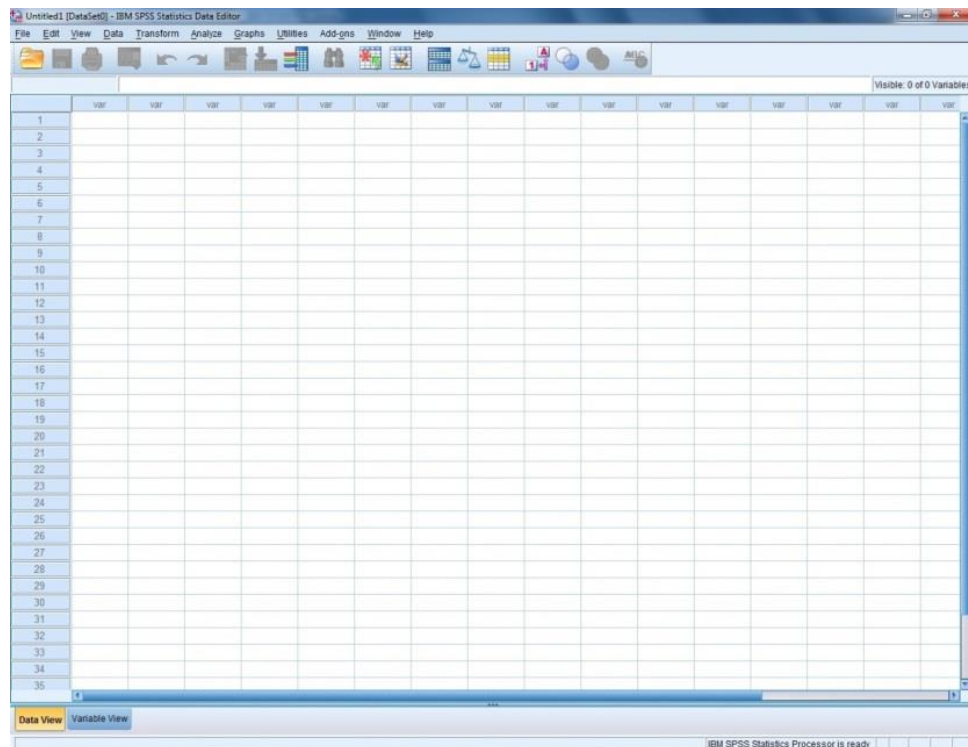
4.4 Relatiotietokanta-malli

Relatiotietokanta koostuu useista tauluista. Tauluissa tieto esitetään riveillä ja sarakkeilla. Relatiotietokannassa pyritään siihen, että yksi tieto tallennetaan vain yhteen paikkaan. Tauluihin tallennetaan tietoa myös siitä, miten eri taulujen tiedot liittyvät toisiinsa.

(Sarja, 14.11.2012)

4.5 SPSS

SPSS (kuva 3) on IBM:n kehittämä tilastotieteelliseen analyysiin suunniteltu ohjelmisto. Nopealla katsauksella SPSS näyttää hyvin paljon Excel-taulukkolaskentaohjelmalta, mutta SPSS poikkeaa Excelistä sen vahvalla painottumisella tilastoanalyysiin. (IBM, 14.11.2012)



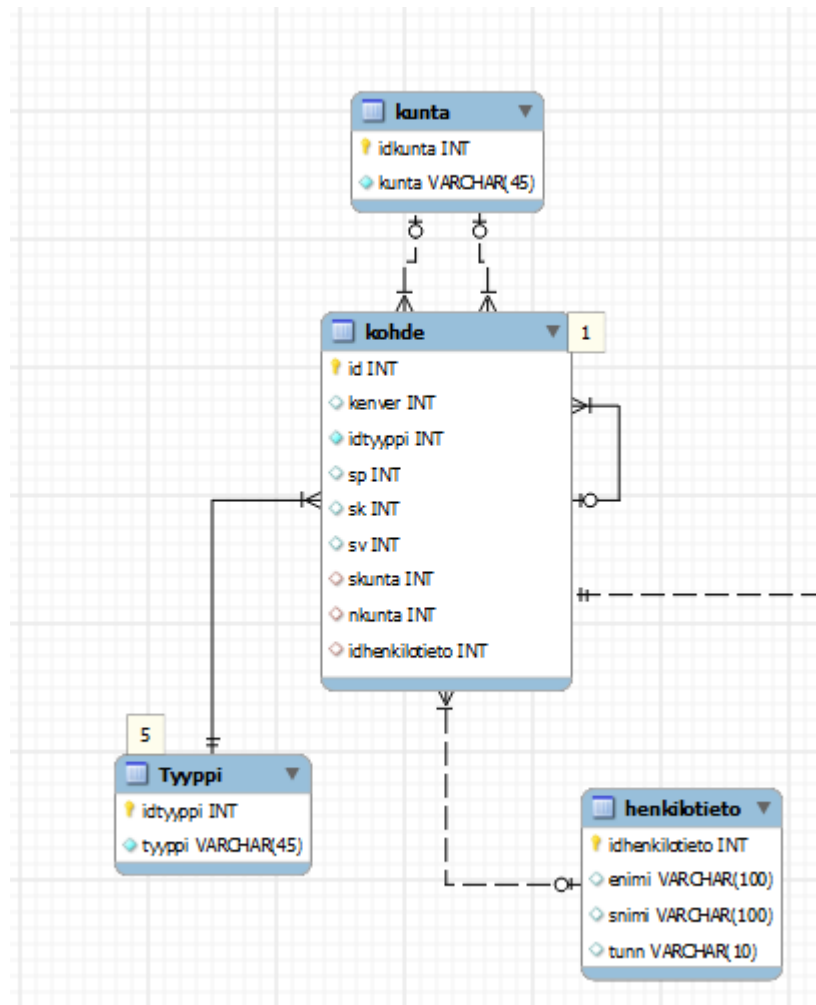
KUVA 3 SPSS-ohjelma

Tässä työssä SPSS-ohjelmaa käytettiin kahteen asiaan. Tutkimusmateriaaliin tutustumiseen ja tutkimus materiaalin siirtämiseen SPSS:n omista sav-tiedostoista csv-tiedostoihin joista siirsin tiedot demo tietokantaan.

Kuvasta näkee tietokannan yleisen rakenteen. Tietokannan eri osat käydään läpi seuraavissa kappaleissa.

5.1 Tietokannan osa: Kohde

Tähän (kuva 5) **Virhe. Viitteen lähdettä ei löytynyt.** tietokannan rakenteeseen tallennetaan tietoja tutkimusprojektin tutkimuskohteista, kuten potilaista, potilaiden verro-
keista ja solulinjoista.



KUVA 5 Kohdekokonaisuus

Kohde-taulu tallennetaan tai oikeastaan kootaan tutkimuskohteiden perustiedoista, kuten syntymäaika, synnyinkunta, nykyinen asuinkunta sekä kenen verrokki kohde mahdollisesti on.

Henkilötieto-tauluun tallennetaan potilaiden ja verrokkien henkilötiedot. Nämä tiedot ovat salassa pidettäviä joten niihin pääsyä on rajattava rajaamalla pääsyä lukemaan tauluntietoja. Näin saadaan eristettyä muut tietokannan tiedot arkaluontoisista henkilö-
tiedoista.

Tyyppi-tauluun tallennetaan tieto tutkimuskohteen tyypistä eli onko kohde potilas, verrokki, solulinja vai benignin eli hyvän laatuksen diagnoosin saanut henkilö. Tämä on pakollinen tieto jokaiselta kohteelta.

5.1.1 Kohde-taulun sisältö

Seuraavassa taulukossa (taulukko 1) näkyy mitä muuttujia kohde-taulurakenteeseen otetaan ja mistä lähteistä.

TAULUKKO 1 Kohde-taulun sisältö

Taulun nimi	Kohde	Henkilötieto	tyyppi
taulun numero workbenchissä	1		5
Huom.			Ryhmä = tyyppi
spss tiedosto	cambridge spss	cambridge spss	cambride
	muuttujat	muuttujat	
	kenver	tunn	ryhma
	verskun	SNIMI	
	skunta	ENIMI	
	nkunta		
	sp		
	sk		
	sv		

Kuten taulukosta näkee niin cambridge numerot spss – tiedostossa olevat muuttujat ovat nimetty samaan tapaan kuin kannan sarakkeet, lukuun ottamatta tyyppi-taulun ryhmä-muuttujaa. Tämä on kannassa nimetty tyyppi. Taulun pääavaimena toimii juokseva integer luku. Kunta tauluun tulee luettelo kunnista joihin verskun, skunta ja nkunta taulujen arvot osoittavat. Käytössäni olleessa materiaalissa tätä kunta listaa ei kuitenkaan ollut. Cambride spss-tiedosto pitää kuitenkin sisällään kunta-muuttujat tutkimuskohteille, kuten yllä olevasta taulukosta käy ilmi (taulukko 1).

5.1.2 Kohde-taulujen väliset suhteet

Seuraavissa taulukoissa (taulukot 2 - 5) käydään läpi Kohde-taulujen väliset suhteet.

TAULUKKO 2 Tyyppi- ja kohde- taulujen suhteet

Taulu	Tyyppi			Kohde
	Avain	Suhde	Ei null	Viiteavain
	idTyyppi	yksi - moneen	Kyllä	idTyyppi

Tyyppi- ja kohde -taulujen välinen suhde on pakollinen. Jokaisella kohteella on oltava tyyppi jota ko. kohde edustaa esimerkiksi syöpäpotilas tai verrokki.

TAULUKKO 3 Kunta- ja kohde- taulujen suhteet

Taulu	Kunta			Kohde
	Avain	Suhde	Ei null	Viiteavain
	idKunta	yksi - moneen	Ei	skunta
	idKunta	yksi - moneen	Ei	nkunta

Kunta- ja kohde- taulujen väliset suhteet eivät ole pakollisia. Silloin kun kohde on solulinja, niin luonnollisesti kohteella ei tällöin voi olla nykyistä kuntaa tai synnyin kuntaa.

TAULUKKO 4 Henkilötieto- ja kohde- taulujen suhteet

Taulu	Henkilötieto			Kohde
	Avain	Suhde	Ei null	Viiteavain
	idhenkilotieto	yksi - moneen	Ei	idhenkilotieto

Henkilötieto- ja kohde- taulujen väliset suhteet eivät ole pakollisia. Silloin kun kohde on solulinja, niin kohteella ei voi olla henkilötietoja.

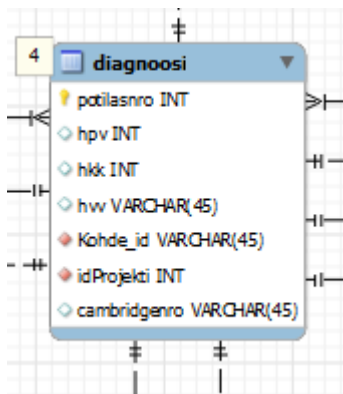
TAULUKKO 5 Kenver- ja id- sarakkeiden suhteet

Taulu	Kohde			Kohde
	Avain	Suhde	Ei null	Viiteavain
	id	yksi - moneen	Ei	kenver

Kenver- ja id- sarakkeiden välinen suhde on kohde-taulun sisäinen suhde. Kohde tauluun on varastoituna terveitä verrokkeja jotka ovat taustoiltaan samantyyppisiä sairastavien potilaiden kanssa, mutta kuitenkin terveitä.

5.2 Tietokannan osa: Diagnoosi

Diagnoosi-taulu (kuva 6) on koko tietokannan keskus. Se pitää sisällään tutkimuskohteina olevien potilaiden, solulinjojen jne. yksilöivät potilasnumerot. Kaikki taulut ovat suorasti tai epäsuorasti yhteydessä tähän tauluun.



KUVA 6 diagnoosi-taulu

Yllä on kuva diagnoosi-taulun rakenteesta (kuva 6). Diagnoosi-taulun sisältö käydään läpi seuraavassa kappaleessa.

5.2.1 Diagnoosi-tilun sisältö

Seuraavassa taulukossa (taulukko 6) näkyy mitä muuttujia kohde-tilurakenteeseen talletetaan ja mistä lähteistä.

TAULUKKO 6 Diagnoosi-tilun sisältö

Taulun nimi	diagnoosi
tilun numero workbenchissä	4
spss tiedosto	cambridge
Huom.	pnro = potilasno
	muuttujat
	hvp
	hkk
	hvv
	Cambridgenro
	pnro

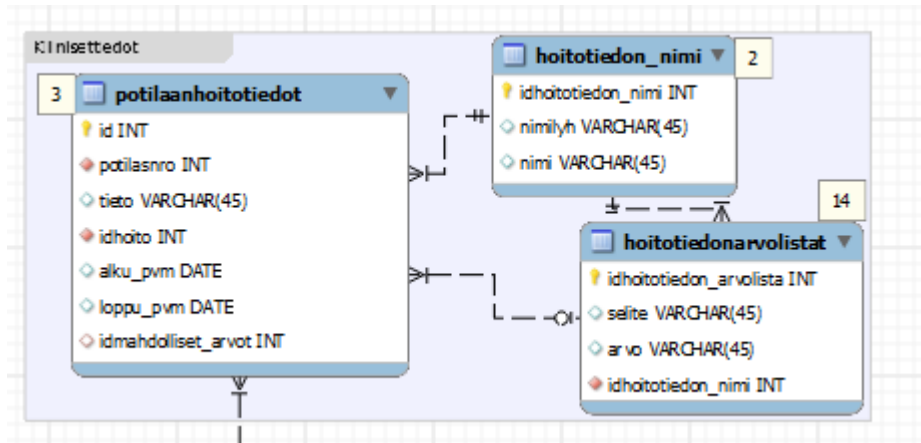
Kuten taulukosta näkee niin cambridge numerot spss –tiedostossa ovat muuttujat ovat nimetty samaan tapaan kuin kannan sarakkeet. Kuitenkin sillä erolla, että lähde materiaalin pnro on nimetty kannassa potilasno:ksi. Cambride numerot spss-tiedostosta tähän tauluun otetaan hvp, hkk, hvv muuttujat jotka kertovat ensimmäisen hoitopäivän potilaalle. Silloin kun kyseessä on verrokki tai solulinja niin silloin luonnollisesti hoitopäivää ei merkata.

Tutkimuskohteet lähdemateriaalissa yksilöidään pnro:n, eli potilasnumeron, avulla joka on juokseva kokonaisluku. Tämä pnro on uudelleen nimetty tähän tietokantaan potilasno:ksi. Potilasno saa arvoiksi jo olemassa olevien kohteiden pnro:t cambridgen numerot spss-tiedostosta. Uudet tutkimuskohteet saavat kannassa juoksevan kokonaisluvun, joka lähtee liikkeelle siitä mihin pnro loppuu.

Cambridgenro on vastaava tutkimuskohteen yksilöintitapa kuin pnro, mutta sitä ei ole annettu kaikille eikä pnro ja cambridgenro ole yhtenevät. Jokaisella tiedolla, jolla on cambridgenro, on myös pnro. Tämä vanha yksilöinti numero kuitenkin kulkeutuu mukana koska sitä on aikaisemmin käytetty tutkimusmateriaalissa.

5.3 Tietokannan osa: Kliiniset tiedot

Tähän (kuva 7) tietokannan rakenteeseen talletetaan kliinisiä tietoja potilaista. Kliiniset tiedot ovat tietoja siitä mitä hoitoja kukin potilas on saanut.



KUVA 7 Kliiniset tiedot -kokonaisuus

Potilaanhoitotiedot-taululla on yhdestä moneen suhde diagnoosi taulun kanssa, joten yhdellä potilasnumerolla voi olla monta eri hoitoa. Tämä rakenne toimii niin, että hoitotiedon_nimi-tauluun on tallennettuna mahdollisten hoitojen nimet ja hoitotiedonarvolistat-tauluun on tallennettu niille mahdolliset arvot. Esimerkiksi muuttuja kumpu, joka kertoo kumman rinnan vuoksi potilas lääkäriin alun perin joutunut. Tallennetaan hoitotiedon_nimi-taulun sarakkeeseen nimilyh. Nimi sarakkeeseen tallennetaan selite lyhenteelle eli "Kumman rinnan vuoksi lääkärissä". Tälle tiedolle mahdolliset arvot tallennetaan hoitotiedonarvolistat-tauluun. Selite kenttään tulee tieto esim. Oikea, vasen, molemmat tai oikea kainalo ja niin edelleen. Näille on spss-tiedostoissa annettu myös integer arvot jotka tallennetaan arvo kenttään.

Rakenteen tietoja voidaan käyttää kahdella eri tavalla. Voidaan käyttää arvolistoja tai tallentaa hoidon arvo suoraan potilaanhoitotiedot-taulun sarakkeeseen tieto ja ohittaa hoitotiedonarvolistat-taulu. Näin täytyy toimia silloin kun hoitotiedolla ei ole ennalta määrättyä arvolistaa josta arvo tälle pystyttäisiin valitsemaan.

Hoitotiedoilla on myös alku- ja loppupäivät jotka ovat merkittynä spss-tiedostoihin muotoon hoidonnimi_al_pv, hoidonnimi_al_kk, hoidonnimi_al_vv jossa hoidonnimi on kunkin hoidon nimi. Esimerkiksi met_hor4 muuttujan alkupäivän muuttujan nimi on muotoa hor4_al_pv. Vastaavasti hoidon loppumis päivämäärä on merkittynä muotoon hoidonnimi_lop_pv, hoidonnimi_lop_kk, hoidonnimi_lop_vv, josta esimerkiksi hor4_lop_kk on

met_hor4 hoidon loppu kuukausi. Nämä hoitojen aloitus- ja lopetuspäivämäärät merkitään potilaanhoitotiedot tauluun sarakkeisiin alku_pvm ja loppu_pvm DATE-muodossa.

5.3.1 Kliiniset tiedot-tilujen sisältö

Seuraavissa taulukoissa (taulukot 7 ja 8) on listattu muuttujat spss-tiedostoista, jotka tallennetaan kliiniset tiedot-tiluihin.

TAULUKKO 7 Potilaanhoitotiedot-tilun sisältö

Potilaanhoitotiedot	Taulun nimi		
4	taulun numero workbenchissä		
maria	spss-tiedosto		
muuttujat			
Chemo_adjuvant_start	hor2_lop_pv	hor8_al_kk	kemo_lop_vv
Chemo_adjuvant_stop	hor2_lop_kk	hor8_al_vv	kemo6_al_pv
Horm_adjuvant_start	hor2_lop_vv	hor8_lop_pv	kemo6_al_kk
Horm_adjuvant_stop	hor3_al_pv	hor8_lop_kk	kemo6_al_vv
u_pr_pv	hor3_al_kk	hor8_lop_vv	kemo6_lop_pv
u_pr_kk	hor3_al_vv	kemo1_al_pv	kemo6_lop_kk
u_pr_vv	hor3_lop_pv	kemo1_al_kk	kemo6_lop_vv
u_pr_h_al_pv	hor3_lop_kk	kemo1_al_vv	kemo7_al_pv
u_pr_h_al_kk	hor3_lop_vv	kemo1_lop_pv	kemo7_al_kk
u_pr_h_al_vv	hor4_al_pv	kemo1_lop_kk	kemo7_al_vv
u_pr_h_lop_pv	hor4_al_kk	kemo1_lop_vv	kemo7_lop_pv
u_pr_h_lop_kk	hor4_al_vv	kemo2_al_pv	kemo7_lop_kk
u_pr_h_lop_vv	hor4_lop_pv	kemo2_al_kk	kemo7_lop_vv
res_pv	hor4_lop_kk	kemo2_al_vv	
res_kk	hor4_lop_vv	kemo2_lop_pv	
res_vv	hor5_al_pv	kemo2_lop_kk	
res_h_al_pv	hor5_al_kk	kemo2_lop_vv	
res_h_al_kk	hor5_al_vv	kemo3_al_pv	
res_h_al_vv	hor5_lop_pv	kemo3_al_kk	
res_h_lop_pv	hor5_lop_kk	kemo3_al_vv	
res_h_lop_kk	hor5_lop_vv	kemo3_lop_pv	
res_h_lop_vv	hor6_al_pv	kemo3_lop_kk	
met_pv	hor6_al_kk	kemo3_lop_vv	
met_kk	hor6_al_vv	kemo4_al_pv	
met_vv	hor6_lop_pv	kemo4_al_kk	
hor1_al_pv	hor6_lop_kk	kemo4_al_vv	
hor1_al_kk	hor6_lop_vv	kemo4_lop_pv	
hor1_al_vv	hor7_al_pv	kemo4_lop_kk	
hor1_lop_pv	hor7_al_kk	kemo4_lop_vv	
hor1_lop_kk	hor7_al_vv	kemo5_al_pv	
hor1_lop_vv	hor7_lop_pv	kemo5_al_kk	
hor2_al_pv	hor7_lop_kk	kemo5_al_vv	
hor2_al_kk	hor7_lop_vv	kemo5_lop_pv	
hor2_al_vv	hor8_al_pv	kemo5_lop_kk	

TAULUKKO 8 Hoitotiedon_nimi-taulun sisältö

hoitotiedon_nimi			Taulun nimi
2			Taulun numero
maria		cambridge	spss-tiedosto
Muuttujat			
FU_DAY	Adj_radiation	met_hor7	kumpu
FU_MONTH	Doses	vaste_hor7	
FU_YEAR	Fields	met_hor8	
FU_DATE	uusi_prim	vaste_hor8	
FU_SURVIVAL	u_pr_op	met_kemo1	
RFS	u_pr_ke	vaste_kemo1	
RFS_DAY	u_pr_sh	met_kemo2	
RFS_MONTH	u_pr_kemo	vaste_kemo2	
RFS_YEAR	u_pr_hor	met_kemo3	
RFS_DATE	lokresid	vaste_kemo3	
RFS_SURVIVAL	res_op	met_kemo4	
OIK_VAS	res_sh	vaste_kemo4	
KAIN_EV	res_kemo	met_kemo5	
LÄHTÖTILANNE	res_hor	vaste_kemo5	
Chemo_adjuvant	kaukomet	met_kemo6	
Chemo_adjuvant_type	met_hor1	vaste_kemo6	
Chemo_adjuvant_dose	vaste_hor1	met_kemo7	
Chemo_anthracycline	met_hor2	vaste_kemo7	
Chemo_taxane	vaste_hor2		
Chemo_CMF	met_hor3		
Chemo_adjuvant_duration	vaste_hor3		
Horm_adjuvant	met_hor4		
Horm_adjuvant_type	vaste_hor4		
Tam_dose	met_hor5		
Horm_adjuvant_duration	vaste_hor5		
Surgery	met_hor6		
Radiation	vaste_hor6		

Kuten taulukosta 8 TAULUKKO 7 Potilaanhoitotiedot-taulun sisältönäkee, niin tauluun potilaanhoitotiedot-tauluun tallennetaan lähdetiedoista hoidon alku- ja loppupäivämäärät. Nämä päivämäärät ovat kuitenkin jaettu päivä, kuukausi ja vuosi tasolle. Joten nämä täytyy kantaan muuttaa date-muotoon.

Potilaanhoitotiedot-tauluun kirjataan hoidon nimestä vain id, joka otetaan hoitotiedon_nimi-taulusta. Kullakin hoidolla voi olla kahdenlaisia mahdollisia arvoja. Hoidolla voi olla joko vapaamuotoinen arvo tai tietty arvo joka on listattuna hoitotiedonarvolistat-taulussa.

Esimerkiksi muuttujalla Chemo_adjuvant_type on kolme mahdollista arvoa 2, 3 ja 5. Nämä arvot tallennetaan arvo-sarakkeeseen hoitotiedonarvolista-aulussa. Näille arvoille on selitteet CMF, CNF ja CEF. Nämä selitteet tallennetaan selite-sarakkeeseen hoitotiedonarvolista-aulussa. Päivämäärät hoidon aloitukselle ja lopetukselle Chemo_adjuvant_start ja Chemo_adjuvant_stop. Poikkeuksellisesti nämä kaksi ovat jo valmiiksi päivämäärä muodossa eikä erikseen päivä, kuukausi ja vuosi tasolla, kuten esimerkiksi res_h_al_pv, res_h_al_kk ja res_h_al_vv. Chemo_adjuvant_dose muuttujalle ei ole ennalta määrättyä arvolistaa. Tämän tyyppisten muuttujien arvot tallennetaan potilaanhoitotiedot-aulun tieto-sarakkeeseen.

Hoitotiedonarvolistat-auluun tallennettavat hoitotietojen muuttujat tulevat spss-tiedostoista. Nämä tiedot on koottuna myös erillisiin asiakkaalle toimitettuihin ohjetiedostoihin cambridge_numerot_selitteet.csv ja Maria_selitteet.csv (6.1 SPSS selite-tiedostot)

5.3.2 Kliiniset tiedot-aulujen väliset suhteet

Seuraavissa taulukoissa (taulukot 9 - 13) käydään läpi kliiniset tiedot-aulujen väliset suhteet.

TAULUKKO 9 Diagnoosi- ja potilaanhoitotiedot – taulujen suhde

Taulu	diagnoosi			potilaanhoitotiedot
	Avain	Suhde	Ei null	Viiteavain
	potilasnro	Yksi -moneen	Kyllä	potilasnumero

Diagnoosi- ja potilaanhoitotiedot – taulujen välinen suhde on yhdestä moneen. Yhdellä potilasnumerolla voi olla useita klinisiä tietoja. Suhde on määriteltä pakolliseksi, mutta esimerkiksi solulinjalla näitä tietoja ei ole.

TAULUKKO 10 Hoitotiedon_nimi- ja potilaanhoitotiedot –taulujen suhde

Taulu	hoitotiedon_nimi			potilaanhoitotiedot
	Avain	Suhde	Ei null	Viiteavain
	idhoitotiedon_nimi	Yksi -moneen	Kyllä	idhoito

Hoitotiedon_nimi- ja potilaanhoitotiedot –taulujen välinen suhde on yhdestä moneen. Sillä sama hoitotieto voi olla merkitty useammalle potilaalle. Tämä suhde on määritetty pakolliseksi, sillä jos potilaanhoitotiedot-auluun merkitään tietoa, niin sillä täytyy olla jokin nimi.

TAULUKKO 11 Hoitotiedon_nimi- ja hoitotiedon_arvolistat –taulujen välinen suhde

Taulu	hoitotiedon_nimi			hoitotiedon_arvolistat
	Avain	Suhde	Ei null	Viiteavain
	idhoitotiedon_nimi	Yksi -moneen	Kyllä	idhoitotiedon_nimi

Hoitotiedon_nimi- ja hoitotiedon_arvolistat –taulujen välinen suhde on yhdestä moneen. Yhdellä hoitotiedolla voi olla monta valmiiksi määriteltä arvoa. Tämä suhde on määriteltä pakolliseksi, koska kaikkien hoitotiedon_arvolistat-aulussa olevat arvot on kuuluttava jollekin hoitotiedolle. Kaikilla hoitotiedoilla ei kuitenkaan ole näitä valmiiksi määriteltä arvoja. Tällöin ko. hoitotiedolle ei merkitä kuuluvaksi yhtäkään arvoa hoitotiedon_arvolistat-aulusta.

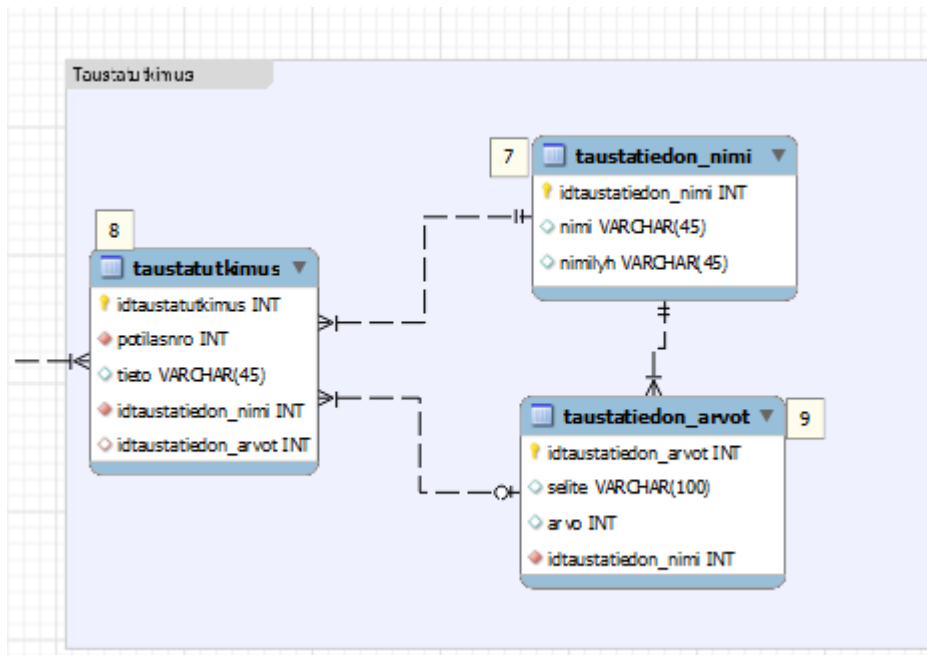
TAULUKKO 12 Hoitotiedon_arvolistat- ja potilaanhoitotiedot –taulujen välinen suhde

Taulu	hoitotiedon_arvolistat			potilaanhoitotiedot
	Avain	Suhde	Ei null	Viiteavain
	idhoitotiedon_arvolista	Yksi -moneen	ei	idmahdolliset_arvot

Hoitotiedon_arvolistat- ja potilaanhoitotiedot –taulujen välinen suhde on yhdestä moneen. Yksi hoitotiedon arvo voi osoittaa moneen eri potilaanhoitotietoon. Taulujen suhde ei ole pakollinen, sillä jos muuttujalla jonka arvo tallennetaan potilaanhoitotiedot-auluun, on vapaamuotoinen, ei tätä suhdetta täytetä ja idmahdolliset_arvot-sarake jää nulliksi potilaanhoitotiedot-aulussa.

5.4 Tietokannan osa: Taustatiedot

Tähän (kuva 8) tietokannan osaan tallennetaan potilailta kerätyt taustatutkimustiedot. Taustatiedot ovat tietoja joilla kartoitetaan potilaan taustaa. Näitä tietoja verrataan kokonaisuutena muuhun tutkimustietoon.



KUVA 8 Taustatutkimuskokonaisuus

Kuten kuvasta yllä näkee, niin kokonaisuus on rakenteeltaan hyvin samanlainen kuin kliiniset tiedot -kokonaisuus (kappale 5.3). Taustatutkimus- ja kliiniset tiedot - kokonaisuudella on kuitenkin selvä ero siinä, että taustatutkimuksilla ei ole päivämäärät niin tärkeästi esillä kuin kliinisissä tiedoissa joissa jokaiselle hoidolle on annettu alku- ja loppupäivät.

Kuten kliiniset tiedot kokonaisuutta, myös taustatutkimus kokonaisuutta voidaan käyttää kahdella tavalla. Rakenteeseen tallennetaan muuttujien arvoja joko arvolistasta tai vapaamuotoisesti taustatutkimus.

5.4.1 Taustatiedon-taulujen sisältö

Seuraavissa taulukoissa (taulukko 13 - 14) listataan ne muuttujat, jotka tallennetaan taustatutkimus-kokonaisuuteen spss-tiedostoista.

Kullekin potilaalle ja verrokille on tehty taustatietokysely. Nämä tiedot on tallennettuna spss-tiedostoon niin, että potilasnumero toimii pääavaimena.

TAULUKKO 13 Taustatiedon_nimi-taulun sisältö 1/2

taustatiedon_nimi		Taulun nimi					
7		Taulun numero					
cam- bride		spss-tiedosto					
amm	rp8	ehika3	lahsika3	vilm1	il6	vaik2	v4
ntyo	rk8	ehkes3	lahsmi4	vitmon1	ml6	vaik3	rask1.ika
nvv	rsp8	ehko4	lahskuk4	vit2	laak6	alrsa	estrogvk
tluo	rpv9	ehika4	lahskum4	vitik2	ikesto7	krinta	vitamin
spit	rp9	ehkes4	lahsika4	vilm2	il7	arsai1	alkok12
spai	rk9	ehko5	lahsmi5	vitmon2	ml7	arsai2	kkika2
kkika	rsp9	ehika5	lahskuk5	vit3	laak7	arsai3	sikä2
rask	rpv10	ehkes5	lahskum5	vitik3	ikesto8	arsai4	kklop2
rpv1	rp10	estro	lahsika5	vilm3	il8	arsai5	kklop3
rp1	rk10	aloiti	lahsmi6	vitmon3	ml8	arsai6	epikä1
rk1	rsp10	lopeti	lahskuk6	pcvit	laak8	toipit	estraika
rsp1	rpv11	kaika	lahskum6	rasvap	ikesto9	omast	kkpitl
rpv2	rp11	mhorm	lahsika6	ikä	il9	okoeky	kkkesto
rp2	rk11	mhko1	tupak	koul	ml9	okyki1	kkestov
rk2	rsp11	mhika1	tupav	opiskv	laak9	okyki2	imetys
rsp2	rpv12	mhkes1	tupafs	imet	ikesto10	okyki3	suksyöpä
rpv3	rp12	mhko2	tupakk	ikesto1	il10	okyki4	suksyöp1
rp3	rk12	mhika2	tupapp	il1	ml10	okyasp	skunta4
rk3	rsp12	mhkes2	tupafv	ml1	laak10	omuu	skunta18
rsp3	rpv13	mhko3	tupafp	laak1	ikesto11	vmast	savonkun
rpv4	rp13	mhika3	pastko	ikesto2	il11	vkoeky	savokun2
rp4	rk13	mhkes3	pastty	il2	ml11	vkyki1	ldcoskun
rk4	rsp13	mhko4	alko	ml2	laak11	vkyki2	neliopit
rsp4	rpv14	mhika4	alkoka	laak2	ikesto12	vkyki3	bmi
rpv5	rp14	mhkes4	alkoik	ikesto3	il12	vkyki4	bmiluokk
rp5	rk14	lahsuk	alkoh	il3	ml12	vkyasp	BCACSTATUS
rk5	rsp14	lahsmi1	alkok	ml3	laak12	vmuu	Contralat_BC
rsp5	klop	lahskuk1	alkoto	laak3	kks	sislkm	AgeatIIBC
rpv6	vkqv	lahskum1	skahvi	ikesto4	vk	velkm	Time- toblooddraw
rp6	ehkp	lahsika1	pkahvi	il4	kkpit	vmitta1	MenopausalS- tatus
rk6	ehko1	lahsmi2	tee	ml4	kkesto	vmitta2	Parity
rsp6	ehika1	lahskuk2	virvj	laak4	loppk	lmita1	
rpv7	eh- kes1	lahskum2	suklaa	ikesto5	lopmi	lmita2	
rp7	ehko2	lahsika2	kaakao	il5	oaik1	dluok1	
rk7	ehika2	lahsmi3	vitam	ml5	oaik2	dluok2	
rsp7	eh- kes2	lahskuk3	vit1	laak5	oaik3	dluok3	
rpv8	ehko3	lahskum3	vitik1	ikesto6	vaik1	tnm	

TAULUKKO 14 Taustatiedon_nimi-taulun sisältö 2/2

Taulun nimi	taustatiedon_nimi
Taulun numero	7
spss-tiedosto	maria
	V2
	TNM_uusi
	T_koodi
	stage_koodi
	histologia_uusi
	histo_koodi
	histo_pelkistetty
	gr_koodi
	N_uusi
	N_koodi
	ER
	ER_koodi
	PR
	PR_koodi
	TRIPLE_NEG

Taustatutkimus-tauluun kirjataan taustatiedon nimestä vain id, joka otetaan taustatiedon_nimi-taulusta. Kullakin taustatiedolla voi olla kahdenlaisia mahdollisia arvoja. Hoidolla voi olla joko vapaamuotoinen arvo tai tietty arvo joka on listattuna taustatiedon_arvot-taulussa.

Esimerkiksi vkkv-muuttujan, eli minkä ikäisenä kuukautiset loppuivat, arvo on vapaamuotoinen koska sillä ei ole arvolistaa. Tämän muuttujan arvo tallennetaan suoraan taustatutkimus-tauluun. Kun taas alrsa-muuttujalla, eli aikaisempi lääkärin toteama rintatasairaus, on arvolista, joten sen arvo valitaan taustatiedon_arvot-taulusta johon nämä arvot ovat valittuina. Tällöin tieto kenttään ei tallenneta mitään. Alrsa-muuttujan ja muut muuttujat tallennetaan taulurakenteeseen seuraavasti. Muuttujan nimi ja nimilyhenne tallennetaan taustatiedon_nimi tauluun. Tässä tapauksessa muuttujan nimi on "aik lääktoteama rintatasairaus" ja nimilyh, eli nimen lyhenne, on "alrsa". Jos tällä muuttujalla ei olisi ennalta määrättyä arvolistaa, josta se voi saada arvoja, niin tämän muuttujan arvo kirjattaisiin taustatutkimus-taulun tieto-sarakkeeseen. Tässä esimerkki tapauksessa kuitenkin muuttujalla kuitenkin löytyy arvolista. Muuttujalla on seuraavat kokonaisluku

arvo ”,00” ja ”1,00”. Näille kahdelle arvolle on selitteet ”Ei” ja ”Kyllä”. Nämä arvot tallennetaan taustatiedon_arvot-tilauun arvo-sarakkeeseen ja selitteet tallennetaan selite-sarakkeeseen.

5.4.2 Taustatiedon-tilujen väliset suhteet

Seuraavissa tilukoissa (tilukot 15 - 18) käydään läpi taustatiedon-tilujen väliset suhteet.

TAULUKKO 15 Diagnoosi- ja taustatutkimus- tilujen välinen suhde

Taulu	diagnoosi			taustatutkimus
	Avain	Suhde	Ei null	Viiteavain
	potilasno	yksi - moneen	Kyllä	potilasno

Diagnoosi- ja taustatutkimus- tilujen välinen suhde on yhdestä moneen. Yhdellä potilasnumerolla voi olla useita taustatutkimustietoja. Suhde on määritelty pakolliseksi, sillä jokaisen taustatutkimustiedon on kuuluttava jollekin potilasnumerolle.

TAULUKKO 16 Taustatiedon_nimi- ja taustatutkimus- tilujen välinen suhde

Taulu	taustatiedon_nimi			taustatutkimus
	Avain	Suhde	Ei null	Viiteavain
	idtaustatiedon_nimi	yksi - moneen	Kyllä	idtaustatiedon_nimi

Taustatiedon_nimi- ja taustatutkimus- tilujen välinen suhde on yhdestä moneen. Yksi taustatiedon_nimi-tilun arvo voi kuulua moneen taustatutkimus-tilun arvoon. Suhde on määritelty pakolliseksi, sillä jokaisella taustatutkimus-tilun arvolla oltava taustatiedon nimi.

TAULUKKO 17 Taustatiedon_nimi- ja taustatiedon_arvot- tilujen välinen suhde

Taulu	taustatiedon_nimi			taustatiedon_arvot
	Avain	Suhde	Ei null	Viiteavain
	idtaustatiedon_nimi	yksi - moneen	Kyllä	idtaustatiedon_nimi

Taustatiedon_nimi- ja taustatiedon_arvot- tilujen välinen suhde on yhdestä moneen. Yhdellä taustatiedon_nimi-tilun arvolla voi olla useita taustatiedon_arvot-tiluun määriteltyjä arvoja. Suhde on määritelty pakolliseksi, sillä jokaisen taustatiedon_arvot-tilun arvon täytyy kuulua jollekin taustatiedon_nimi-tilun arvolle.

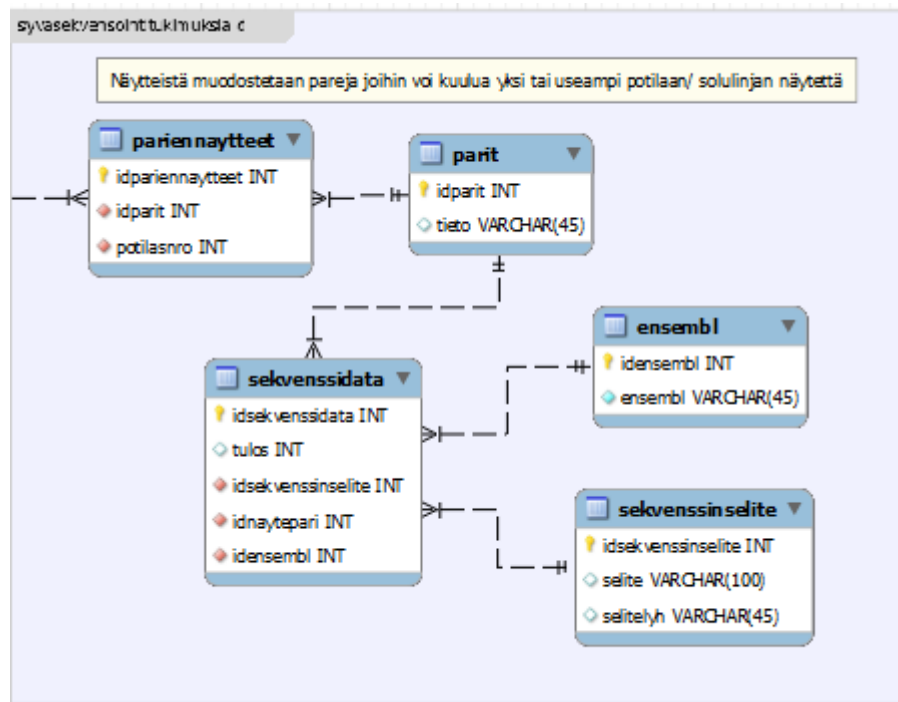
TAULUKKO 18 Taustatiedon_arvot- ja taustatutkimus- taulujen välinen suhde

Taulu	taustatiedon_arvot			taustatutkimus
	Avain	Suhde	Ei null	Viiteavain
	idtaustatiedon_arvot	yksi - moneen	Ei	idtaustatiedon_arvot

Taustatiedon_arvot- ja taustatutkimus- taulujen välinen suhde on yhdestä moneen. Yksi taustatiedon_arvot-taulun arvo voi liittyä moneen taustatutkimus-taulun arvoon. Taulujen suhde ei ole pakollinen. Vapaamuotoisella muuttujalla jonka arvo tallennetaan taustatutkimus-tilaan, ei tätä suhdetta täytetä ja idtaustatiedon_arvot-sarake jää nulliksi taustatutkimus-tilaan.

5.5 Tietokannan osa: Syväsekvenssi

Tähän (kuva 9) tietokannan rakenteeseen kootaan syväsekvenssintutkimuksista saaduista tuloksista kertynyt data.



KUVA 9 Syväsekvenssikokonaisuus

Syväsekvenssintutkimuksissa tutkitaan tiettyjen proteiinien esiintymistä veri- tai kudoksenäytteessä. Tätä dataa tuotetaan kolmessa eri muodossa. Raaka, normalisoitua ja laskettua dataa. Raaka data on nimensä mukaan raakaa eli prosessoimatonta dataa. Tässä datassa käy ilmi tiettyjen proteiinien ilmentyminen kappaleittain näytteessä. Normalisoitua data on normalisoitu jotta siitä saadaan vertailu kelpoista. Laskettu data pitää

sisällään erilaisia tilastollisia laskutoimituksia jotka on tehty normalisoidun datan pohjalta.

5.5.1 Syväsekvenssi-taulujen sisältö

Sisältö näihin syväsekvenssi-tauluihin tulee joko csv tai xls muodossa. Kaikille kolmelle erilaiselle tietotyyppille on oma tiedostonsa. Saman tutkimuksen tiedostot on eriteltynä samaan kansioon.

Seuraavassa on esitettyä esimerkkitapaus siitä kuinka tieto siirtyy näistä lähdetiedostoista tietokantaan. Kaikki tiedot on värikoodattu, jotta tietojen liikkumista olisi helpompi seurata.

Lähdetiedostoissa olevat tiedot on esitettyä seuraavissa (taulukot 19 - 20) kolmessa taulukoissa.

TAULUKKO 19 Normalisoitu data

Tiedosto: NormData.xls				
	Treated_MCF_D APT	Cont- rol_MCF_DMSO	Treated_MDA_D APT	Cont- rol_MDA_DMSO
ENSG000000000 003	583.864614075 464	700.0067060125 19	73.43862871140 61	19.01285425154 89
ENSG000000000 005	0	0.390410878980 769	0	0.864220647797 676
ENSG000000000 419	10960.1384751 387	12810.16176111 7	5769.894423352 91	4729.879605396 68
ENSG000000000 457	2568.71454281 836	2737.170672534 17	4680.223959499 88	4637.407996082 33
ENSG000000000 460	2009.47945340 613	2686.807669145 65	5400.716451992 87	6199.918927300 53
ENSG000000000 938	0	4.684930547769 23	0	34.56882591190 7

TAULUKKO 20 Raaka data

Tiedosto: RawData.xls				
	Treated_MCF_D APT	Cont- rol_MCF_DMSO	Treated_MDA_D APT	Cont- rol_MDA_DMSO
ENSG000000000 003	403	1793	37	22
ENSG000000000 005	0	1	0	1
ENSG000000000 419	7565	32812	2907	5473

TAULUKKO 21 Laskettu data

Tiedosto: annotatedData_Treated_VS_Control.xls				
Feature identifier	Mean expression...	Mean expression ...	Standard deviation...	Standard deviation...
ID	mean_Treated	mean_Control	sd_Treated	sd_Control
ENSG00000158578	2	29	0	0
ENSG00000088926	31	2	1	0

Taulukoissa 19 ja 20 tiedoston nimi on korostettu keltaisella kun taas taulukossa 21 tämä on tehtynä "Feature identifier"-riville. Tämä keltainen korosteväri kertoo selitteen datalle, joka on kaikissa taulukoissa korostettuna punertavalla värillä. Taulussa 21 on tämän lisäksi vielä sininen korosteväri joka on tämän keltaisella korostetun selitteen lyhenne. Oranssi korosteväri kertoo sarakkeen, jossa on lueteltuna ensembl-koodit. Näitä voidaan ajatella tässä yhteydessä tutkittavien proteiinien niminä, jolloin punertava korosteväri kertoo raakadatatassa tämän proteiinin esiintymis määrän näytteessä. Vihreä korosteväri taulukoissa 19 ja 20 kertoo näytteen nimen jolle tämä tutkimus on tehty. Tässä esimerkki tapauksessa nämä ovat solulinjoja joita on kasvatettu koetta varten. Tässä otoksessa on kaksi eri solulinjaa MCF_DAPT ja MCF_DMSO. Näille kahdelle solulinjalle on kummallekin Treated ja Control parit eli lääkkeellä hoidettu ja kontrolli, johon lääke aineen tehoa verrataan.

Seuraavissa (taulukot 22 - 29) taulukoissa on esitetty taulukkojen TAULUKKO 19, TAULUKKO 20 ja

TAULUKKO 21 tiedot tietokannan tauluihin lisättyinä. Huomaa, että osa tauluista on kauempaa tietokannasta kuin syväsekvenssi kokonaisuudesta.

TAULUKKO 22 Kohde-taulu

id	nimi	kenver	idtyyppi
0	Treated_MCF_DAPT	2	1
1	Control_MCF_DMSO	1	1
2	Treated_MDA_DAPT	4	1
3	Control_MDA_DMSO	3	1

TAULUKKO 23 Kohteen tyyppi-taulu

idtyyppi	tyyppi
0	potilas
1	solulinja
2	verkkokki

Vihreällä korostetut näytteiden nimet tallennetaan kohde kokonaisuuden tauluun kohde ja näille lisätään tarvittava tyyppi tyyppi-taulusta. Tässä tapauksessa kaikki ovat solulinjoja.

TAULUKKO 24 Diagnoosi-taulu

kohde_id	pnro
0	1
1	2
2	3
3	4

Kohde kokonaisuudesta näytteiden id siirtyy diagnoosi-tauluun josta se luetaan eteenpäin.

TAULUKKO 25 Pariennaytteet-taulu

idpariennaytteet	idparit	pnro
0	0	1
1	1	2
2	2	1
3	2	2
4	2	3
5	2	4

Pariennaytteet-taulu muodostaa monesta moneen suhteen diagnoosi-taulun pnro-sarakkeen ja parit-taulun idparit-sarakkeen kanssa. Näin voidaan muodostaa sellaisia pareja tutkimuskohteista jotka voivat kuulua useaan eri pariin.

TAULUKKO 26 Parit-taulu

idparit	tieto
0	
1	
2	annotatedData_Treated_VS_Control

Kullekin parille voidaan määritellä lisätieto tieto-sarakkeeseen.

TAULUKKO 27 Sekvenssi-taulu

idensembl	arvo	selite	idparit
0	583.864614075464	1	0
0	700.006706012519	1	1
0	403	0	1
2	10960.1384751387	1	1
1	2	2	2
1	29	3	2
346	0.000283942790480830	4	2

Tutkimuksen datat kerätään selitteineen ja paritietoineen sekvenssi-tauluun

TAULUKKO 28 Syväsekvenssin selite-tila

idselite	selite	selitelyh
0	raaka	
1	normalisoitu	
2	Mean expression	mean_Treated
3	Mean expression	mean_Control

Syväsekvenssin selite-tilaan kootaan selitteet ja niiden lyhenteet.

TAULUKKO 29 Ensembl-tila

idensembl	ensembl
0	ENSG000000000003
1	ENSG00000158578
2	ENSG000000000419
...	...
345	ENSG00000001497
346	ENSG00000158578

Ensembl-tilasta löytyy kaikki tutkimuksissa esiintyneet ensembl koodit.

5.5.2 Syväsekvenssi-tilojen väliset suhteet

Seuraavissa taulukoissa (taulukot 30 - 34) käydään läpi syväsekvenssi-tilojen väliset suhteet.

TAULUKKO 30 Diagnoosi- ja pariennaytteet- tilojen välinen suhde

Taalu	Diagnoosi			pariennaytteet
	Avain	Suhde	Ei null	Viiteavain
	potilasno	yksi - moneen	Kyllä	potilasno

Diagnoosi- ja pariennaytteet- tilojen välinen suhde on yhdestä moneen. Yksi potilasnumero voi olla liitettyä moneen pariin. Suhde on määritelty pakolliseksi, sillä jokaisessa parissa täytyy olla jokin potilasnumero

TAULUKKO 31 Parit- ja pariennaytteet- tilojen välinen suhde

Taalu	parit			pariennaytteet
	Avain	Suhde	Ei null	Viiteavain
	idparit	yksi - moneen	Kyllä	idparit

Parit- ja pariennäytteet- taulujen välinen suhde on yhdestä moneen. Yhteen pariin voi kuulua monta näyte paria. Suhde on määritelty pakolliseksi, sillä jokaisessa parissa täytyy olla näytteitä.

TAULUKKO 32 Parit- ja sekvenssidata- taulujen välinen suhde

Taulu	parit			sekvenssidata
	Avain	Suhde	Ei null	Viiteavain
	idparit	yksi - moneen	Kyllä	idnaytepari

Parit- ja sekvenssidata- taulujen välinen suhde on yhdestä moneen. Yhdestä parista voi olla monta riviä dataa. Suhde on määritelty pakolliseksi, sillä jokaisella data rivillä täytyy olla jokin näyte pari jolle ko. tieto kuuluu.

TAULUKKO 33 Ensembl- ja serkvenssidata- taulujen välinen suhde

Taulu	ensembl			sekvenssidata
	Avain	Suhde	Ei null	Viiteavain
	idensembl	yksi - moneen	Kyllä	idensembl

Ensembl- ja serkvenssidata- taulujen välinen suhde on yhdestä moneen. Yksi ensembl tieto voi viitata moneen eri sekvenssidata- taulun riviin, mutta yksi sekvenssi- taulun rivi ei voi kuulua montaa eri ensembliä. Suhde on määritelty pakolliseksi, sillä jokaisella sekvenssidata- taulun rivillä täytyy olla yksi ensembl.

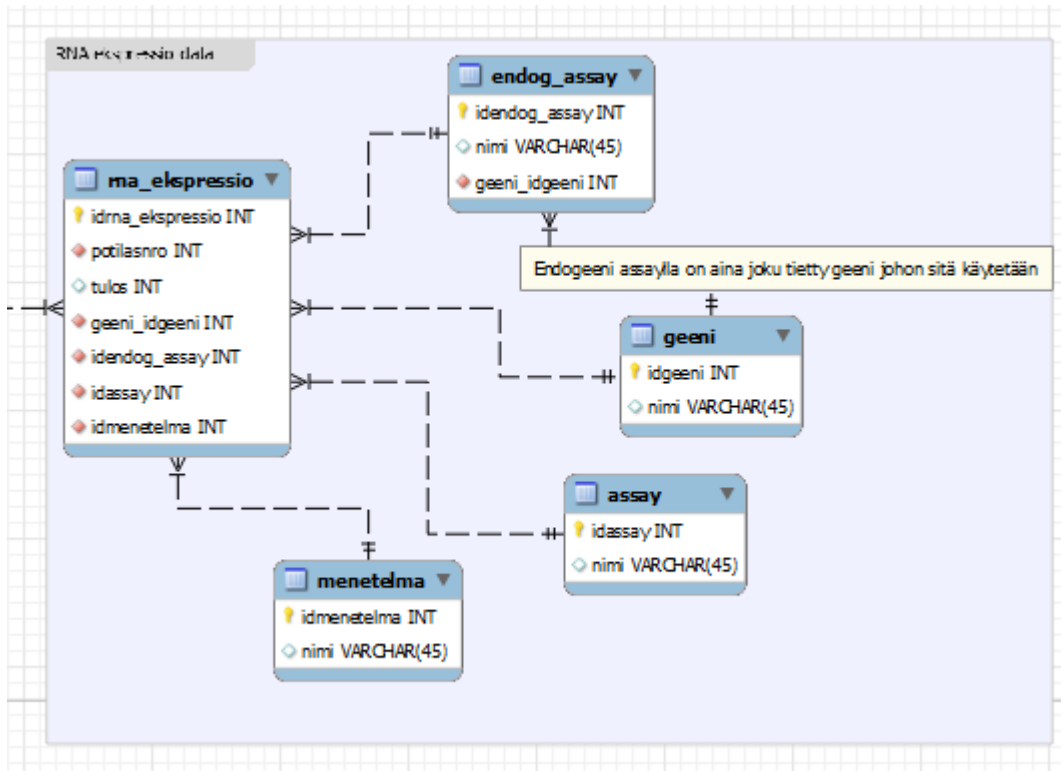
TAULUKKO 34 Sekvenssinselite- ja sekvenssidata- taulujen välinen suhde

Taulu	sekvenssinselite			sekvenssidata
	Avain	Suhde	Ei null	Viiteavain
	idsekvenssinselite	yksi - moneen	Kyllä	idsekvenssinselite

Sekvenssinselite- ja sekvenssidata- taulujen välinen suhde on yhdestä moneen. Yksi selite voi kuulua moneen sekvenssidatan riviin. Suhde on määritelty pakolliseksi, sillä jokaisella sekvenssidatan rivillä täytyy olla selite.

5.6 Tietokannan osa: RNA-ekspressio

Tähän (kuva 12) tietokannan osaan kootaan RNA-ekspressio tutkimuksista saatavat tulokset



KUVA 10 RNA-ekspressio -kokonaisuus

Potilaasta otettuja kudoksia tutkitaan RNA-ekspressiossa siten, että näytteestä valmistetaan ensin kontrolli näyte tietyllä endogeeni assayllä. Tämän jälkeen samasta kohdasta otettua näytettä käsitellään toisella assayllä. Näitä kahta vertaamalla saadaan sitten tietty tulos.

5.6.1 RNA-ekspressio -taulujen sisältö

Seuraavissa taulukoissa (taulukot 35 - 40) käydään läpi RNA-ekspressio -taulujen sisältöä. Tulokset on jaettu eri spss-tiedostoihin tutkittavan geenin perusteella. Nämä tiedostot ovat tutkijoiden käsin kokoamia. Seuraavassa on esitettyä esimerkkitapaus siitä kuinka tieto siirtyy näistä lähdetiedostoista tietokantaan. Kaikki tiedot on värikoodattu, jotta tietojen liikkumista olisi helpompi seurata.

TAULUKKO 35 RNA-ekspressio esimerkki lähdetiedosto

pnro	Aph1a	AssayGeeni	Kontrolli	AssayKontrolli	Menetelmä
34	2,448995363	Hs00211268_m1	PPIA	Hs99999904_m1	Taqman
44	1,453706388	Hs00211268_m1	PPIA	Hs99999904_m1	Taqman
55	1,750655931	Hs00211268_m1	PPIA	Hs99999904_m1	Taqman

Tässä taulukossa (taulukko 35) on ote yhdestä Rna-ekspression lähdetiedosta. Kaikissa tiedostoissa ei ole saraketta *Kontrolli* ja sarake *AssayKontrolli* saattaa olla nimeltään *indendog_assay*. Oletuksena kuitenkin on tämän hetkessä materiaalissa, että jokaisessa näissä tutkimuksissa on käytetty kontrolli geeninä PPIA nimistä geeniä. Lisäksi sarake *AssayGeeni* saattaa olla nimetty eri tavoin. Sarakkeiden järjestys on kuitenkin aina sama. Keltainen korostusväri merkkää geenin nimeä. Vaalean vihreällä korostettuun sarakkeeseen on merkitty käytetty assay. Oranssilla korostetut sarakkeet kertovat käytetyn endogeeni assayn, ensimmäinen sarake, ja siinä käytetyn geenin nimen, toinen sarake. Tumman vihreällä korostettu sarake kertoo käytetyn menetelmän. Punertavalla korostetussa sarakkeessa on tutkimuksen tulos.

TAULUKKO 36 Rna_ekspressio-taulu

idrna_ekspressio	potilasno	tulos	idgeeni	idendog_assay	idassay	idmenetelma
0	34	0.983743	0	0	0	0
1	34	2.448995363214840	2	0	2	0
2	34	0.892	3	0	1	1

RNA-ekspression tulokset kerätään rna_ekspressio-tauluun yllä esitetyllä tavalla.

TAULUKKO 37 Endog_assay-taulu

idendog_assay	nimi	geeni
0	Hs99999904_m1	1

Endog_assay tauluun tallennetaan kaikki käytettävät endogeeni assayt ja niissä käytetyt geenit.

TAULUKKO 38 Geeni-taulu

idgeeni	nimi
0	BRG1
1	PPIA
2	Aph1a
3	RNF4

Geeni-tauluun tallennetaan kaikki geenit.

TAULUKKO 39 Assay-taulu

idassay	nimi
0	Hs00231324_m1
1	Hs00231302_m1
2	Hs00211268_m1

Assay-tauluun tallennetaan kaikki käytetyt assayt.

TAULUKKO 40 Menetelmä-taulu

idmenetelma	nimi
0	Taqman
1	Sequenom

Menetelmä-tauluun tallennetaan kaikki käytetyt menetelmät.

5.6.2 RNA-ekspressio –taulujen väliset suhteet

Seuraavissa taulukoissa (taulukot 41 - 46) käydään läpi RNA-ekspressio –taulujen väliset suhteet.

TAULUKKO 41 Diagnoosi- ja rna_ekspressio- taulujen välinen suhde

Taulu	diagnoosi			rna_ekspressio
	Avain	Suhde	Ei null	Viiteavain
	potilasno	yksi - moneen	Kyllä	potilasno

Diagnoosi- ja rna_ekspressio- taulujen välinen suhde on yhdestä moneen. Yhdellä potilaalla voi olla useita RNA-ekspressio tuloksia. Suhde on määritelty pakolliseksi, sillä jokaisella RNA-ekspressio tuloksella täytyy olla jokin potilas jolle se kuuluu.

TAULUKKO 42 RNA_ekspressio- ja endog_assay- taulujen välinen suhde

Taulu	rna_ekspressio			endog_assay
	Avain	Suhde	Ei null	Viiteavain
	idendog_assay	yksi - moneen	Kyllä	idendog_assay

RNA_ekspressio- ja endog_assay- taulujen välinen suhde on yhdestä moneen. Yhdenlaista endogeeni assaytä voidaan käyttää useassa tutkimuksessa. Suhde on määriteltä pakolliseksi, sillä jokaisella tutkimuksella täytyy olla merkittynä mitä endogeeni assaytä siinä on käytetty.

TAULUKKO 43 Geeni- ja endog_assay- taulujen välinen suhde

Taulu	geeni			endog_assay
	Avain	Suhde	Ei null	Viiteavain
	idgeeni	yksi - moneen	Kyllä	geeni_idgeeni

Geeni- ja endog_assay- taulujen välinen suhde on yhdestä moneen. Sama geeni on voinut olla monen eri endogeeni assayn pohjana. Suhde on määriteltä pakolliseksi, sillä jokaisella endogeeni assayllä täytyy olla merkittynä se mitä geeniä siinä on käytetty.

TAULUKKO 44 Geeni- ja rna_ekspressio- taulujen välinen suhde

Taulu	geeni			rna_ekspressio
	Avain	Suhde	Ei null	Viiteavain
	geeni_idgeeni	yksi - moneen	Kyllä	geeni_idgeeni

Geeni- ja rna_ekspressio- taulujen välinen suhde on yhdestä moneen. Samaa geeniä voidaan tutkia monessa eri tutkimuksessa. Suhde on määriteltä pakolliseksi, sillä jokaisessa tutkimuksessa täytyy olla merkittynä tutkittava geeni.

TAULUKKO 45 Assay- ja rna_ekspressio- taulujen välinen suhde

Taulu	assay			rna_ekspressio
	Avain	Suhde	Ei null	Viiteavain
	idassay	yksi - moneen	Kyllä	idassay

Assay- ja rna_ekspressio- taulujen välinen suhde on yhdestä moneen. Samanlaista assaytä voidaan käyttää useassa eri tutkimuksessa. Suhde on määriteltä pakolliseksi, sillä jokaisessa tutkimuksessa on merkittävä mitä assaytä käytettiin.

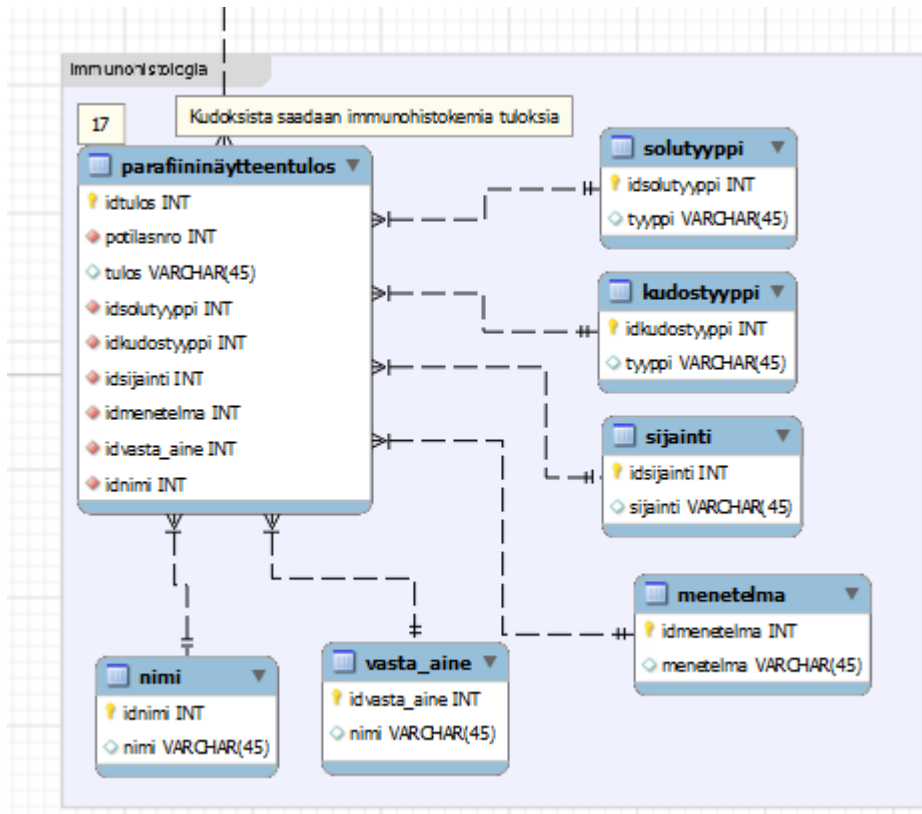
TAULUKKO 46 Menetelmä- ja rna_ekspressio- taulujen välinen suhde

Taulu	menetelmä			rna_ekspressio
	Avain	Suhde	Ei null	Viiteavain
	idmenetelmä	yksi - moneen	Kyllä	idmenetelmä

Menetelmä- ja rna_ekspresso- taulujen välinen suhde on yhdestä moneen. Samaa menetelmää voidaan käyttää monessa eri tutkimuksessa. Suhde on määritelty pakolliseksi, sillä jokaisessa tutkimuksessa täytyy merkitä käytetty menetelmä.

5.7 Tietokannan osa: Immunohistologia

Tähän (kuva 11) tietokannan rakenteeseen kootaan kaikki immunohistologia tulokset.



KUVA 11 Immunohistologiakokonaisuus

Immunohistologiassa tutkitaan parafiiniin säilöttyjä kudoksenäytteitä värjäämällä niitä tietynlaisilla proteiineilla ja tutkimalla mikroskoopilla värjäytyneitä kohtia. Kappaleessa 3 on kerrottu lisää tästä tutkimus menetelmästä. Kuten yllä olevasta kuvasta näkee, niin jokaisella tuloksella on monta määrittelevää tekijää.

5.7.1 Immunohistologia-taulujen sisältö

Seuraavissa taulukoissa (taulukot 47 - 54) käydään läpi immunohistologia-taulujen sisältöä.

TAULUKKO 47 Ote immunohistologian yhdestä lähdetiedostosta

pnro	PSEN1_membr_IHC	Kudostyyppi	Solutyyppi	Sijainti	Vasta_aine	Menetelmä
34	0	carcinoma	epiteeli	membraani	Chemicon MAB5232	IHC
42	0	carcinoma	epiteeli	membraani	Chemicon MAB5232	IHC

Yllä olevassa taulukossa on ote *PSEN1_IHC_membr.sav* tiedostosta, joka on yksi immunohistologian lähdetiedostoista. Taulukoissa on käytetty korostevärejä, jotta tiedon liikettä lähdetiedostosta tietokantaan olisi helpompi seurata. Varsinainen tutkimustulos on korostettu punertavalla värillä. Tumman sininen korosteväri on tutkittavan proteiinin nimi. Oranssilla korosteella on merkitty tutkittavaa kudostyyppiä. Keltaisella korosteella on merkitty solutyyppiä. Vaalean vihreällä korosteella on merkitty löydetyn proteiinin sijainti. Vaalean sinisellä korosteella on merkitty tutkimuksessa käytetty vasta-aine. Tumman vihreällä korosteella merkitään käytettyä menetelmää.

TAULUKKO 48 Parafiininäytteenäytös-taulu

id_tulos	potilasno	tulos	idsolutyyppi	idkudostyyppi	idsijainti	idmenetelmä	idvasta_aine	idnimi
0	34	0	0	0	0	0	0	0
1	34	2	0	0	1	0	0	1
2	43	3	0	0	1	0	0	1

Parafiininäytteenäytös-taulu on se taulu johon tallennetaan nämä immunohistologian tulokset. Taulu on hyvin samankaltainen kuin lähdemateriaalin taulukko.

TAULUKKO 49 Solutyyppi-taulu

idsolutyyppi	tyyppi
0	epiteeli

Solutyyppi-tauluun tallennetaan kaikki immunohistologia tutkimuksissa esiintyneet solutypit.

TAULUKKO 50 Kudostyyppi-taulu

idkudostyyppi	tyyppi
0	carcinoma

Kudostyyppi-tauluun tallennetaan kaikki immunohistologia tutkimuksissa esiintyneet kudostypit.

TAULUKKO 51 Sijainti-taulu

idsijainti	sijainti
0	membraani
1	sytosoli

Sijainti-tauluun tallennetaan kaikki immunohistologia tutkimuksissa havaitut proteiinin esiintymis kohdat.

TAULUKKO 52 Menetelmä-taulu

idmenetelma	menetelma
0	IHC

Menetelmä tauluun tallennetaan kaikki immunohistologia tutkimuksissa käytetyt menetelmät.

TAULUKKO 53 Vasta_aine-taulu

idvasta_aine	nimi
0	Chemicon MAB5232

Vasta-aine-tauluun tallennetaan kaikki immunohistologia tutkimuksissa käytetyt vastaaineet.

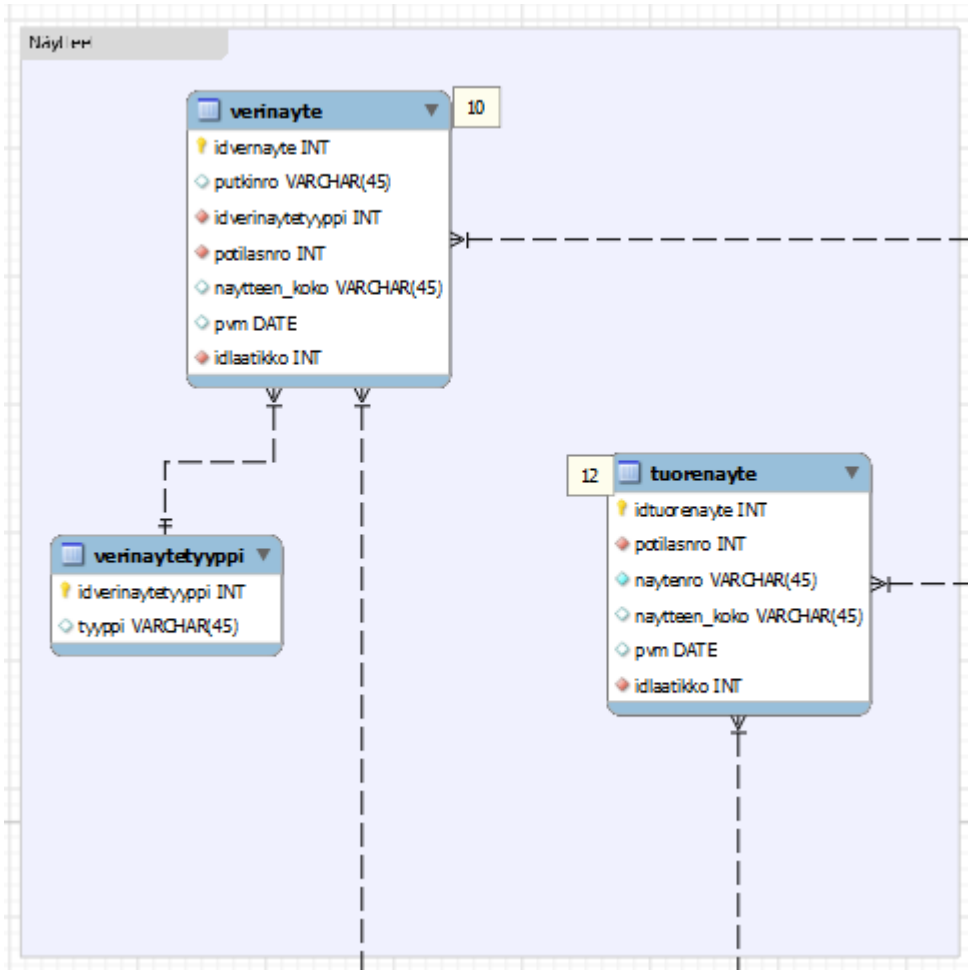
TAULUKKO 54 Nimi-taulu

idnimi	nimi
0	PSEN1_membr_IHC
1	PSEN1

Nimi-tauluun tallennetaan kaikki immunohistologia tutkimuksissa esiintyneet proteiinien nimet.

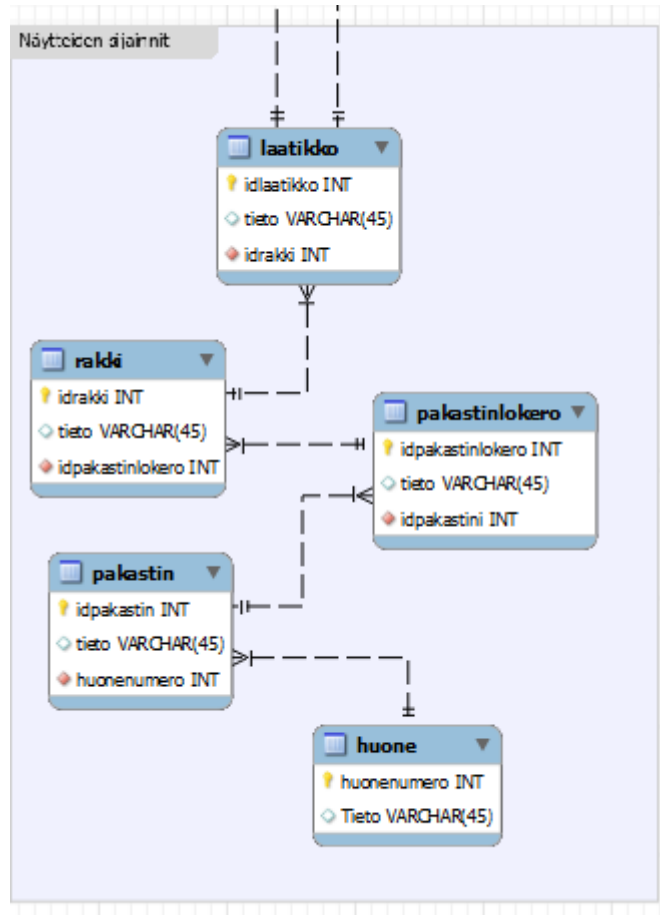
5.8 Tietokannan osa: Näytteet ja näytteiden sijainti

Tähän tietokannan osaan tallennetaan tiedot otetuista veri- ja kudospnäytteistä (kuva 12Kuva 12), sekä tiedot niiden sijainnista (kuva 13).



KUVA 12 Näytteetkokonaisuus

Potilaista otettujen verinäytteiden tyyppi, putkinumero, koko ja päivämäärä tallennetaan verinäyte-tauluun. Vastaavasti kudospnäytteistä tallennetaan koko, näytenumero ja päivämäärä.



KUVA 13 Näytteiden sijainti -kokonaisuus

Tuore- ja verinäytteiden sijaintitiedot tallennetaan samaan taulu rakenteeseen (kuva 13KUVA 13). Näitä näytteitä säilytetään erilaisissa pakastimissa joissa on lokeroita. Näissä lokeroissa on räkkejä joihin on laitettu laatikoita ja nämä laatikot pitävät sisäl-
lään itse näytteet.

5.8.1 Näytteet-tilujen sisältö

Seuraavassa käydään läpi veri- ja tuoreinäytteiden tietojen ja sijainnin tallennus tietokantaan. Koska tutkimusprojektissa näytteiden sijainnin tiedot on tallennettu siten, että näytteiden putkinumerot, tyypit ja säilytys laatikot on kirjattu uusiverit231111 Excel-tiedostoon ja tämän jälkeen kun tiedetään onko näyte "normaalissa" pakastimessa vai -80 °C pakastimessa voidaan näytteen laatikko numeroa lähteä etsimään pakastimen kylkeen laitetusta lapusta, jossa lukee siihen talletetut laatikot. Tietokannan suunnittelussa on kuitenkin otettu huomioon asiakkaan toive tarkemmat järjestelmän varalle ja tietokannassa on taulut valmiina laatikosta aina huoneen tasolle asti. Tällä nykyisellä materiaalilla kuitenkin on sijaintitiedoista vain laatikkoon numero digitoituna, joten tietokantaan voidaan tallentaa vain tälle tasolle menevät sijaintitiedot.

Tuoreinäytteiden tiedoista on digitoituna vain potilas, josta näyte on otettu, näytteen numero ja päivämäärä, jona näyte on otettu. Näytteiden sijainnista ei ole digitoitua tietoa.

5.8.1.1 Verinäytteet

Verinäytteiden tiedot on tallennettuna uusiverit231111.xlsx-tiedostoon. Alla (TAULUKKO 55 ja TAULUKKO 56) on ote tästä tiedostosta. Nämä otteet on jaettu kahteen osaan, jotta se mahtuisi tähän A4:llä helposti luettavassa muodossa. Tiedosto on käsin laadittu ja täten ulkopuoliselle hieman työläs lukea. Väri koodasin taulukosta samaa tarkoittavat sarakkeet, jotta sitä olisi helpompi lukea.

TAULUKKO 55 uusiverit231111.xlsx osa 1/2

PUTKI	ML	PNRO	PV	"-80°C/LTK"	"-80°C/PU"	EDTA/LTK	ISO SE/LTK
2451	11	1232	9.8.1993				26
3084		1675	6.2.1995	63	4007	9	41

TAULUKKO 56 uusiverit231111.xlsx osa 2/2

ISO SE/PU	PIE SE/LTK	PIE SE/PU	HEPVE/LTK	HEPVE/ML	HEP/LTK	HEP/PU	ML
2566	33	2566	4	3	24	2566	0
4007	51	4007	10	2	38	4007	6

Keltainen korosteväri kertoo putkinumeron kunkin tyyppiselle verinäytteelle. Verinäytteen tyyppi on kerrottuna näiden sarakkeiden otsikossa. Paitsi ensimmäisessä sarakkeessa, joka kuuluu EDTA-verelle. Vaalean sininen korosteväri kertoo sen vasemman puoleisen putken näytteen koon millilitroina. Tätä näytteen koko tietoa ei ole tallennettu seeruminäytteille ja -80 °C pakkaseen tallennetulle näytteelle. Lisäksi Hepariiniverelle ja hepariiniseerumille on sama putkinumero, jota ei ole erikseen merkitty. Tämä putkinumero on merkittynä vain *HEP/PU*-sarakkeeseen. Tällöin hepariiniveren koon kertovan sarakkeen, *HEPVE/ML*, vasemmalla puolella ei ole putkinumeroa, kuten EDTA ja hepariiniverellä. Oranssi korosteväri kertoo potilaan potilasnumeron jolta näyte on otettu. Vihreä korosteväri kertoo päivämäärän jolloin näyte on potilaasta otettu. Huomaa, että usealla putkella on sama putkinumero.

TAULUKKO 57 Verinäyte-taulu

idverinäyte	putkinro	idverinäytteentyppi	potilasno	näytteen_koko	pvm	idlaatikko
0	3084	0	1675		6.2.1995	0
1	4007	1	1675		6.2.1995	1
2	4007	2	1675		6.2.1995	2
3	4007	3	1675		6.2.1995	3
4	4007	4	1675	6	6.2.1995	5
5	4007	5	1675	2	6.2.1995	4

Verinäyte-tauluun tallennetaan näytteiden putkinumerot, potilasnumerot, verinäytteen tyypit, näytteen otto päivämäärä ja laatikon numero, josta näyte löytyy.

TAULUKKO 58 Verinäytteentyppi-taulu

idverinäytteentyppi	tyyppi
0	EDTA
1	-80°C
2	iso seerumi
3	pieni seerumi
4	Hepariiniveri seerumi
5	Hepariiniveri

Verinäytteentyppi-tauluun tallennetaan verinäytteen tyyppi.

TAULUKKO 59 Laatikko-taulu

idlaatikko	laatikkonro	idrakki
0	9	
1	63	
2	41	
3	51	
4	10	
5	38	

Laatikko-tauluun tallennetaan jokaisen laatikon numero. Jatkossa kun näytteiden sijaintitietoja digitoidaan enemmän, niin tähän laitetaan myös räkin numero josta se löytyy.

5.8.1.2 Tuorenäyte

Tuorenäytteiden tiedot on tallennettuna tuorenäyte xls-tiedostoon. Tiedosto on käsin laadittu ja täten ulkopuoliselle hieman työläs lukea. Väri koodasin taulukon (TAULUKKO 60) sarakkeita ja karsin tuorenäytteitä käsittelemättömät sarakkeet pois tästä esimerkistä.

TAULUKKO 60 Ote tuorenäyte.xls-tiedostosta

PNRO	PV	TUORENÄYTE
1442	21.4.1994	
1096	19.3.1993	N6758/A

Tuorenäyte xls-tiedostossa on tallennettuna tiedot tuorenäytteen potilaasta, josta näyte on otettu, päivämäärä, jolloin näyte otettiin ja itse tuorenäytteen numero. Nämä tiedot tallentuvat tuorenäyte-tauluun kuten esimerkissä alla (TAULUKKO 61).

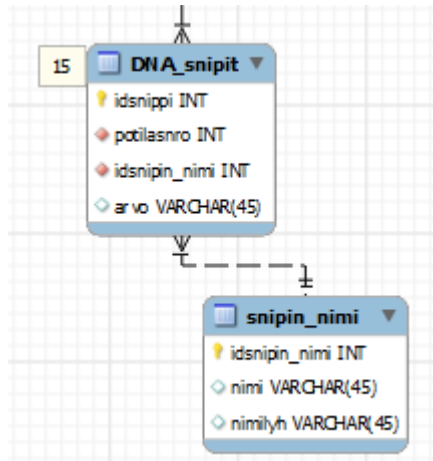
TAULUKKO 61 Tuorenäyte-taulu

idtuorenäyte	potilasno	naytenro	naytteen_koko	pvm	idlaatikko
0	2082	N6758/A		19.3.1993	

Tuorenäyte-taulussa on sarake näytteen koolle tulevaisuutta varten sekä suhde näytteen sijainti kokonaisuuteen. Tuorenäytteiden sijaintitietoja ei kuitenkaan ole vielä digitoitu, joten niitä ei tämän työn teko hetkellä tähän tietokantaan voida tallentaa.

5.9 Tietokannan osa: DNA snipit

Tähän (kuva 14) taulu rakanenteeseen tallennetaan potilaiden DNA snipi tietoja. DNA snipi tiedot ovat dna:n emäs pareja tietyssä kohtaa dna:ta.



KUVA 14 DNA snipit -kokonaisuus

Tähän tauluun voi tulevaisuudessa tulla hyvinkin monta riviä tietoa per potilas. Tällä hetkelle kuitenkin potilaista on 26 eri snipiä valittu tallennettavaksi tähän tietokantaan olemassa olevasta lähdemateriaalista.

5.9.1 DNA snipit taulujen sisältö

DNA snipi tiedot on tallennettuna cambridge numerot –spss tiedostoon. Tähän tauluun tallennettavat sarakkeet on eriteltyä alla olevaan taulukkoon (taulukko 62).

TAULUKKO 62 Snipin_nimi-taulun sisältö

Nimi	Nimilyh
D22S426	geno426
RS738977	genosnp1
S3827351	genosnp2
RS3484	genosnp3
RS760517	genosnp4
D22S1177	geno1177
RS738148	genosnp5
RS7285064	genosnp6
RS25095	genosnp7
RS733655	genosnp8
RS228941	genosnp9
D22S445	geno445
RS228942	snp10
rs5756515	genot515
rs2543519	genot519
rs2235324	genot324
rs2235321	genot321
rs732756	genotyypp
rs2235325	genot325
rs1421312	ge312
rs5756505	genot505
rs855788	genot788
rs4820268	rs268ge
rs2246092	rs092ge
rs2017764	rs764ge

Nimi sarakkeessa on varsinainen snipin nimi ja Nimilyh sarakkeessa on lähdemateriaalissa enemmän käytetty nimi.

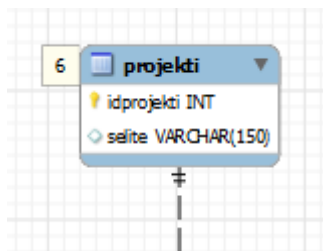
TAULUKKO 63 Ote Cambridge numerot spss -tiedostosta

pnro	genosnp1	ge-not505
726	CC	GC
732	TT	GG

Yllä (taulukko 63) olevassa taulukossa on ote Cambridge numerot spss –tiedostosta. Otteesta käy ilmi muoto jossa tiedot ko. tiedostossa on esitetty. Otin otteeseen kahden eri potilaan genosnp1 ja genot505 muuttujat.

5.10 Tietokannan osa: Projekti

Tähän (kuva 15) tietokannan osaan on tarkoitus tallentaa projektin nimi johon kukin potilas kuuluu.



KUVA 15 Projekti-taulu

Tällä hetkellä yksi potilas voi olla osa yhtä projektia kerrallaan, mutta tarpeen vaatiessa tämä on helppo muuta vaihtamalla projekti- ja diagnoosi-taulujen suhde monesta monen suhteeksi.

6 Ohjetiedostot

Tässä kappaleessa käydään läpi asiakkaalle toimitettujen ohjetiedostojen käyttöä joita tarvitaan suunnitellun kannan käytössä ja jatkokehityksessä. Tiedostoja on kahdenlaisia karttoja ja selitteitä.

Kartta-tiedostoissa kerrotaan mistä lähteestä tieto siirretään mihinkin kohtaa tietokantaa. Selite-tiedostossa kerrotaan spss-tiedostojen muuttujien määränpää tietokannassa, sekä muuttujien selitteet ja arvolistat.

6.1 SPSS selite-tiedostot

Näitä selite tiedostoja on kaksi. Yksi kummallekin spss-tiedostolle. Nämä spss-tiedostot ovat toimineet tiedontallennus paikkana tutkimustiedolle ennen tätä tietokantaa. Nämä spss-tiedostot ovat: cambridge_numerot_32010.sav (lyh. cambridge) ja Maria_030311.sav (lyh. maria). Näistä tiedostoista on myös csv versiot.

6.1.1 SPSS selite-tiedostojen rakenne

Seuraavassa taulukossa (taulukko 64) on ote cambridge_numerot_selitteet-tiedostosta.

TAULUKKO 64 Ote cambridge_numerot_selitteet-tiedostosta

nimilyh	pnro	her2
taulunro	1	17
nimi	POTILASNUMERO	HER2-analyysi
arvot		0 negative
		1 faintly positive
		2 moderate
		3 strongly positive

Cambridge_numerot_selitteet-tiedostossa on annettuna kaikki cambride spss tiedostosta tietokantaan talletettavat muuttujat. Joukossa on muutama muuttuja joita ei oteta mukaan tietokantaan. Näitä muuttujia merkitään taulunumerolla nolla(0).

Nimilyh-sarake kertoo muuttujan nimi lyhenteen. Hoitotiedon_nimi ja taustatiedon_nimi -tauluissa on saman nimiset sarakkeet joihin tämä tieto tallennetaan.

Taulunro-sarake kertoo taulun numeron joka on annettuna joillekin tärkeille tauluille workbenchissä. Tämä on tarkoitus helpottaa taulujen löytämistä.

Nimi-sarake kertoo muuttujan koko nimen tai selitteen. Hoitotiedon_nimi ja taustatiedon_nimi –tauluissa on samannimiset sarakkeet joihin tämä tieto tallennetaan.

Arvot-sarake kertoo onko muuttujalla tietty arvolista josta arvo valitaan vai onko se vapaamuotoinen. Esimerkiksi pnro:lla ei ole listattuja arvoja joten se on vapaamuotoinen. Kun taas muuttujalla her2 on listattuna 4 erilaista arvoa. Täten sillä on ennalta määrätty arvolista, joka tallennetaan tauluun 17 eli taustatiedon-arvot. Nämä arvot muodostuvat kahdesta osasta. Arvosta ja sen selitteestä. Arvo on numeraalinen arvo ja selite tekstiä. Jokaisen arvon muoto selite-tiedostossa on muotoa. arvo selite.

7 Yhteenveto

Tässä työssä toteutettiin suunnitelma tietokannasta, johon tulisi tallennettavaksi Itä-Suomen yliopiston patologian yksikön rintasyövän tutkimuksessa syntyvää materiaalia. Materiaali tulisi olla helposti saatavilla kaikille yksikön tutkijoille ja tietokanta poistaisi rinnakkaisten aineistojen käytöstä nykyisin aiheutuvat ongelmat.

Tätä suunnittelu projektia tehtiin Itä-Suomen yliopiston tiloissa. Työaikaa meni noin kolme kuukautta suunnitelman tekoon josta noin kuukausi meni tutkimusprojektiin perehtymiseen ja tarpeen kartoittamiseen. Työn aikana toteutettiin demo tietokanta malliksi siitä kuinka tiedot sinne tallentuisivat ja kuinka niihin pääsisi käsiksi. Työn tuloksena tuotettiin tämän raportin lisäksi MySQL Workbench:llä toteutettu suunnitelma tietokannasta ja tutkimusaineiston siirtoa tietokantaan kuvaavia dokumentteja, jotka eivät ole sisällytettynä kokonaisuudessaan tähän dokumenttiin. Nämä tiedostot on luovutettu asiakkaalle.

Projektin tuotokset antavat mielestäni hyvän mahdollisuuden jatkaa projektia toteutusasteelle. Omasta mielestäni paras ratkaisu olisi, että projektia jatkettaisiin räätälöidyllä ohjelmalla joka toteuttaisi kaksi asiaa. Tiedon lataamisen ja päivittämisen sekä tiedon ulos saannin tietyin parametrein csv-tiedostoina. Näin itse syöväntutkimusprosessi ei muuttuisi ja tutkijat voisivat jatkaa työtä jo olemassa olevilla tutkimusprosesseilla.

Lähteet

Itä-Suomen yliopisto - Kliininen patologia ja oikeuslääketiede. [Viitattu 12.11.2012].

Saatavissa: <http://www.uef.fi/kliinisenlaaketieteenyksikko/kliininen-patologia-ja-oikeuslaaketiede>

Daniel J. Power. A Brief History of Spreadsheets. [Viitattu 14.11.2012]. Saatavissa:

<http://dssresources.com/history/sshistory.html>

Microsoft. Business Intelligence | Database Management | Data Warehousing | Microsoft SQL Server. [Viitattu 14.11.2012]. Saatavissa:

<http://www.microsoft.com/sql-server/en/us/default.aspx>

IBM. IBM SPSS software. [Viitattu 14.11.2012]. Saatavissa: [http://www-](http://www-01.ibm.com/software/analytics/spss/)

[01.ibm.com/software/analytics/spss/](http://www-01.ibm.com/software/analytics/spss/)

Oracle. MySQL :: MySQL Workbench 5.2. [Viitattu 14.11.2012]. Saatavissa:

<http://www.mysql.com/products/workbench/>

Jari Sarja. Relaatiotietokanta. [Viitattu 14.11.2012]. Saatavissa: [http://www.verkkope-](http://www.verkkopedagogi.net/vanhat/fi/sisalto/materiaalit/access2003/luku0375c6.html?C:D=419702&selres=419702)

[dagogi.net/vanhat/fi/sisalto/materiaalit/ac-](http://www.verkkopedagogi.net/vanhat/fi/sisalto/materiaalit/access2003/luku0375c6.html?C:D=419702&selres=419702)

[cess2003/luku0375c6.html?C:D=419702&selres=419702](http://www.verkkopedagogi.net/vanhat/fi/sisalto/materiaalit/access2003/luku0375c6.html?C:D=419702&selres=419702)

Arto Mannermaan haastattelu [02.11.2012] Itä-Suomen Yliopistolla.

Aula Perttu, Kääriäinen Helena, Palotie Aarno. Painos 2006. Perinnöllisyyslääketiede. Julkaisia Duodecim.